DATA MINING AND METHODS FOR EARLY DETECTION, HORIZON SCANNING, MODELLING, AND RISK ASSESSMENT OF INVASIVE SPECIES

EDITED BY: Aristides Moustakas and Stelios Katsanevakis
PUBLISHED IN: Frontiers in Applied Mathematics and Statistics, Frontiers in Ecology and Evolution, Frontiers in ICT, Frontiers in Earth Science, Frontiers in Marine Science and Frontiers in Environmental Science

frontiers Research Topics

ISSN 1664-8714
ISBN 978-2-88945-471-6
DOI 10.3389/978-2-88945-471-6

About Frontiers

Frontiers is more than just an open-access publisher of scholarly articles: it is a pioneering approach to the world of academia, radically improving the way scholarly research is managed. The grand vision of Frontiers is a world where all people have an equal opportunity to seek, share and generate knowledge. Frontiers provides immediate and permanent online open access to all its publications, but this alone is not enough to realize our grand goals.

Frontiers Journal Series

The Frontiers Journal Series is a multi-tier and interdisciplinary set of open-access, online journals, promising a paradigm shift from the current review, selection and dissemination processes in academic publishing. All Frontiers journals are driven by researchers for researchers; therefore, they constitute a service to the scholarly community. At the same time, the Frontiers Journal Series operates on a revolutionary invention, the tiered publishing system, initially addressing specific communities of scholars, and gradually climbing up to broader public understanding, thus serving the interests of the lay society, too.

Dedication to quality

Each Frontiers article is a landmark of the highest quality, thanks to genuinely collaborative interactions between authors and review editors, who include some of the world's best academicians. Research must be certified by peers before entering a stream of knowledge that may eventually reach the public - and shape society; therefore, Frontiers only applies the most rigorous and unbiased reviews.

Frontiers revolutionizes research publishing by freely delivering the most outstanding research, evaluated with no bias from both the academic and social point of view.

By applying the most advanced information technologies, Frontiers is catapulting scholarly publishing into a new generation.

What are Frontiers Research Topics?

Frontiers Research Topics are very popular trademarks of the Frontiers Journals Series: they are collections of at least ten articles, all centered on a particular subject. With their unique mix of varied contributions from Original Research to Review Articles, Frontiers Research Topics unify the most influential researchers, the latest key findings and historical advances in a hot research area! Find out more on how to host your own Frontiers Research Topic or contribute to one as an author by contacting the Frontiers Editorial Office: **researchtopics@frontiersin.org**

DATA MINING AND METHODS FOR EARLY DETECTION, HORIZON SCANNING, MODELLING, AND RISK ASSESSMENT OF INVASIVE SPECIES

Topic Editors:
Aristides Moustakas, Universiti Brunei Darussalam, Brunei
Stelios Katsanevakis, University of the Aegean, Greece

'Invasive rabbitfish in eastern Mediterranean', photo by Dimitris Poursanidis.

Invasive alien species are non-indigenous taxa introduced to areas beyond their natural distribution and bio-geographical barriers by human activity, with important impacts on biodiversity, human health and ecosystem services. With the human population being higher than ever before and increasing, together with unprecedented rates of mobility of humans and goods, the introduction of new invasive species is more common than ever and is at the forefront of research in many disciplines such as ecology, epidemiology and food security.

The mechanisms of successful introduction, establishment and spread of invasive alien species are highly complex as biological, social, geographic, economic and climatic factors influence the way an invasive species is introduced and determine the options available for its eventual detection and control.

With the rapid development of smart sensors, social networks, digital maps and remotely-sensed imagery, spatio-temporal data are more ubiquitous and richer than ever before. The availability of such large datasets (Big data) poses great challenges in data analysis. In addition, increased availability of computing power facilitates the use of computationally-intensive methods for the analysis of such data.

Thus new methods are needed to efficiently study and understand biological invasions. A Research Topic held in Frontiers Environmental Informatics aimed to address this topic. Methods are defined in the widest terms and may be analytical, practical or conceptual. Among others, a key aim of the thematic was to maximize the use of the proposed methods/techniques by the scientific community and environmental stakeholders.

Citation: Moustakas, A., Katsanevakis, S., eds. (2018). Data Mining and Methods for Early Detection, Horizon Scanning, Modelling, and Risk Assessment of Invasive Species. Lausanne: Frontiers Media. doi: 10.3389/978-2-88945-471-6

Table of Contents

frontiers
in Applied Mathematics and Statistics

EDITORIAL
published: 21 March 2018
doi: 10.3389/fams.2018.00005

Editorial: Data Mining and Methods for Early Detection, Horizon Scanning, Modelling, and Risk Assessment of Invasive Species

Aristides Moustakas[1]* and *Stelios Katsanevakis*[2]*

[1] *Institute for Applied Data Analytics, Universiti Brunei Darussalam, Bandar Seri Begawan, Brunei,* [2] *Department of Marine Sciences, University of the Aegean, Mytilene, Greece*

Keywords: data mining, alien species, risk assessment, data analytics, ecology, biological invasions, computational modeling, methods development

Editorial on the Research Topic

Data Mining and Methods for Early Detection, Horizon Scanning, Modelling, and Risk Assessment of Invasive Species

Alien species are taxa introduced to areas beyond their natural distribution by human activities, overcoming bio-geographical barriers [1]. Accelerating movement of humans, animals and goods are driving an increasing rate of biological invasions [2, 3]. Through a multitude of mechanisms of introduction, broadly classified into five categories—release, escape, contaminant, stowaway, and corridor—many thousands of species have moved into new regions [2]. In Europe alone, ~15,000 alien species have been recorded, with an increasing trend in new introductions [4]. For most taxonomic groups, the global rate of new introductions is increasing with no sign of saturation in the accumulation of alien species [5].

Many alien species have become invasive with substantial impacts on biological diversity, human health and ecosystem services. Impacts of invasive species range from reduction in individual fitness of native species to population declines, local extinctions, changes in community composition, effects on entire ecosystem processes, and wider ecosystem function, health problems in humans or even deaths, and substantial economic losses [6–11]. The ecological impacts of invasive species are so severe that they are considered as one of the major drivers of global biodiversity loss [12, 13].

The Convention on Biological Biodiversity (CBD) recognized the need for the "compilation and dissemination of information on alien species that threaten ecosystems, habitats, or species to be used in the context of any prevention, introduction and mitigation activities," and calls for "further research on the impact of alien invasive species on biological diversity" [14]. Indeed, research efforts to better understand the factors controlling alien species introduction, initial dispersal, establishment success, distribution, abundance, spatio-temporal dynamics, and invasiveness have been intensified. In the 1990s and 2000s, research on human-mediated invasions has exploded, with an exponential increase of publications dealing with or mentioning biological invasions (**Figure 1**). In addition, many peer-reviewed journals specifically targeting biological invasions have been established (e.g., "Biological Invasions," "Aquatic Invasions," "Management of Biological Invasions," "BioInvasions Records"), a large number of dedicated information systems and online databases have been created [15], and a number of international initiatives advancing research and collaboration in the field have been promoted [16].

Edited and reviewed by:
Xuan Zhu,
Monash University, Australia

****Correspondence:***
Aristides Moustakas
arismoustakas@gmail.com
Stelios Katsanevakis
stelios@katsanevakis.com

Specialty section:
This article was submitted to Environmental Informatics, a section of the journal Frontiers in Applied Mathematics and Statistics

Received: *05 February 2018*
Accepted: *05 March 2018*
Published: *21 March 2018*

Citation:
Moustakas A and Katsanevakis S (2018) Editorial: Data Mining and Methods for Early Detection, Horizon Scanning, Modelling, and Risk Assessment of Invasive Species. Front. Appl. Math. Stat. 4:5. doi: 10.3389/fams.2018.00005

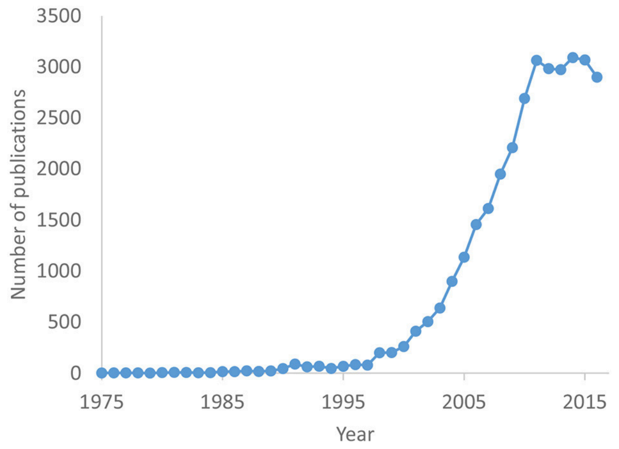


FIGURE 1 | Number of publications registered in Scopus by year, including the words "biological invasions" or "non-indigenous species" or "alien species" or "invasive species."

Despite the intensification of research efforts, there are still substantial gaps in our knowledge on the dynamics of biological invasions and the accumulation of alien species across regions and taxa. Understanding the spread and successful introduction of invasive species is highly complex as biological, social, geographic, economic, and climatic factors influence the way an invasive species is introduced and determine the options available for its eventual detection and control. This in turn calls for an interdisciplinary approach as traditional ecological approaches need to be combined with data mining and data analytics [17]. Data availability has often been the limiting factor in predictive modeling [18, 19]. With the rapid development of smart sensors, social networks, digital maps, and remotely-sensed imagery, spatio-temporal data are more ubiquitous and richer than ever before [20]. The availability of such large datasets (big data) poses great challenges in data analysis [21]. In addition, increased availability of computing power facilitates the use of computationally-intensive methods for the analysis of such data [22]. Thus new methods are needed to understand biological invasions and to efficiently use new technologies and handle big datasets. The current Research Topic in Environmental Informatics aims to address this interesting and cutting-edge theme. Specifically, the special issue focuses on data analysis and novel methods for early detection, horizon scanning, modeling, and risk assessment of invasive species. These themes have been rapidly developing or are expected to develop in the near future (**Figure 2**). Ten contributions were finally accepted after the peer review process.

Data in alien species inventories are often scattered in numerous disconnected formats and databases that lack interoperability entangling their analysis. In their contribution, Groom et al. describe seven ways that data on alien species can be made more accessible and useful. These include data management plans, interoperability of information sources, metadata documentation, existing data formulation, use of common data semantics, data availability, and long-term data preservation.

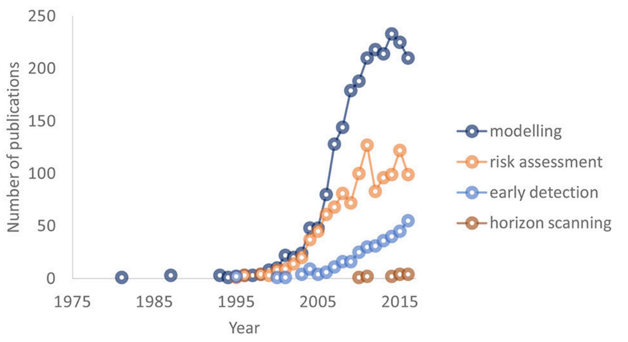


FIGURE 2 | Number of publications registered in Scopus by year, including the words "biological invasions" or "non-indigenous species" or "alien species" or "invasive species" and one of the following words: "modeling," "risk assessment," "early detection," and "horizon scanning."

Spatially explicit high resolution data are elementary for assessing the impacts of alien species and mitigation; however, often these data are unavailable. Environmental explanatory covariates available at the resolution of the alien species, as well as at finer resolutions, can be used to infer alien species presences at finer resolutions using random forests, a machine learning technique, as suggested by Daliakopoulos et al.

Building and managing large datasets of alien species is elementary for their monitoring and management. The geodatabase of the European Alien Species Information Network (EASIN), an initiative of the European Commission, is presented by Deriu et al. The key feature of EASIN is that it aggregates, integrates, and harmonizes spatio-temporal data of alien species in Europe, through standardized processes, supporting both research and policies on biological invasions.

Land use change is the major contributor to the introduction and spread of alien species, and forest lodging and road constructions increase the exposure to infectious diseases. In addition, climatic changes can facilitate alien species and disease spread. The spatio-temporal spread of Hemorrhagic fever with renal syndrome (HFRS) is studied using powerful statistical methods and core areas in China, and the association of climatic cycles are quantified by He et al. demonstrating the applicability of modern space-time modeling and mapping techniques to assess biological invasions.

Islands are undergoing strong pressure from alien species due to their restricted size and increased human impact. Hypotheses of invasibility were explicitly tested by Bjarnason et al. using a detailed statistical framework, and it was concluded that areas of high species richness have greater invasibility. A negative relationship between alien and endemic species richness was also recorded potentially providing inverse learning patterns between alien and endemic species.

The fact that closely related species share a higher chance of becoming infected or attacked by pests allows the identification of taxa with different degrees of vulnerability. Using publicly available information about pest-host interactions and their geographic distributions, in combination with host phylogenetic reconstructions, Robles-Fernández and Lira-Noriega estimated a pest-host interaction or infection index in geographic space as spatially explicit tool for risk assessment.

Species distribution models are a commonly used tool for the analysis of alien species. However, they strongly rely on species' presence/absence and confirmed absences are hard to quantify. Environmental DNA is a novel technique that presents higher detectability and accuracy in comparison to conventional sampling techniques, and can effectively differentiate between species' presence/absence, improving species distribution models, as argued by Muha et al.

Calibrated Individual based models (IBMs) are useful tools for investigating the dynamics of alien species. Samson et al. highlight how early-stage data-limited IBMs can play an important role in defining key research priorities of invasive species in order that subsequent models can provide robust insight into potential management interventions.

Ensemble methods are meta-algorithms that combine several techniques into a unique predictive model to minimize variance; they are increasingly employed in science including a recent application to draw up the habitats' maps for mosquito alien species [23]. However, the development of these models should not be done in a "black box" mode and it should be accompanied by a set of in-depth analysis regarding key training and operation decision points, thereby promoting a reproducible evaluation of the proposed methodology, as suggested by Demertzis et al.

Uncertainty is an attribute of (lack of) information and it is a vital component in other scientific disciplines; however, it is usually missing from marine invasion science studies. The final contribution in this Research Topic, by Katsanevakis and Moustakas, argues that without assessing uncertainty, it is hard to envisage future improvements. In addition, if scientists are perceived by the public either to overstate their findings in order to receive high visibility or to downplay the uncertainty of their findings, society is likely to lose confidence in the outputs of invasion science.

We believe that this Research Topic provides useful new insights for advancing the field of invasion science, by combining knowledge and techniques from multiple disciplines. Science initially went from the phase of "homo-universalis" to high specialization potentially missing the inter-connection with other scientific disciplines. While this may have been a necessity in the history of science, in the big data and globalization era the necessity and feasibility of a more integrated data-based but combining practical and theoretical knowledge approach appears to be more feasible than ever before.

AUTHOR CONTRIBUTIONS

All authors listed have made a substantial, direct and intellectual contribution to the work, and approved it for publication.

ACKNOWLEDGMENTS

We thank all contributors for submitting their work in this Research Topic, the reviewers for their time and effort, the editorial office of Frontiers for coordination and support, and the Associate Editors Xuan Zhu, Yang Liu, and Juergen Piltz for handling the editorial as well as other contributions that AM and SK were coauthoring.

REFERENCES

1. Blackburn TM, Pyšek P, Bacher S, Carlton JT, Duncan RP, Jarošík V, et al. A proposed unified framework for biological invasions. *Trends Ecol Evol.* (2011) **26**:333–9. doi: 10.1016/j.tree.2011.03.023
2. Essl F, Bacher S, Blackburn TM, Booy O, Brundu G, Brunel S, et al. Crossing frontiers in tackling pathways of biological invasions. *Bioscience* (2015) **65**:769–82. doi: 10.1093/biosci/biv082
3. Canning-Clode J. *Biological Invasions in Changing Ecosystems: Vectors, Ecological Impacts, Management and Predictions.* Warsaw; Berlin: Walter de Gruyter GmbH & Co KG (2015).
4. Katsanevakis S, Deriu I, D'Amico F, Nunes AL, Sanchez SP, Crocetta F, et al. European alien species information network (EASIN): supporting European policies and scientific research. *Manage Biol Invas.* (2015) **6**:147–57. doi: 10.3391/mbi.2015.6.2.05
5. Seebens H, Blackburn TM, Dyer EE, Genovesi P, Hulme PE, Jeschke JM, et al. No saturation in the accumulation of alien species worldwide. *Nat Commun.* (2017) **8**:14435. doi: 10.1038/ncomms14435
6. Vilà M, Basnou C, Pyšek P, Josefsson M, Genovesi P, Gollasch S, et al. How well do we understand the impacts of alien species on ecosystem services? A pan-European, cross-taxa assessment. *Front Ecol Environ.* (2010) **8:**135–44. doi: 10.2307/20696461
7. Blackburn TM, Essl F, Evans T, Hulme PE, Jeschke JM, Kühn I, et al. A unified classification of alien species based on the magnitude of their environmental impacts. *PLOS Biol.* (2014) **12**:e1001850. doi: 10.1371/journal.pbio.1001850
8. Katsanevakis S, Wallentinus I, Zenetos A, Leppäkoski E, Çinar ME, Oztürk B, et al. Impacts of marine invasive alien species on ecosystem services and biodiversity: a pan-European review. *Aquat Invas.* (2014) **9**:391–423. doi: 10.3391/ai.2014.9.4.01
9. Mazza G, Tricarico E, Genovesi P, Gherardi F. Biological invaders are threats to human health: an overview. (2014) *Ethol Ecol Evol.* **26**:112–29. doi: 10.1080/03949370.2013.863225
10. Bellard C, Cassey P, Blackburn TM. Alien species as a driver of recent extinctions. *Biol Lett.* (2016) **12**:20150623. doi: 10.1098/rsbl.2015.0623
11. Ojaveer H, Galil BS, Campbell ML, Carlton JT, Canning-Clode J, Cook EJ, et al. Classification of non-indigenous species based on their impacts: considerations for application in marine management. *PLoS Biol.* (2015) **13**:e1002130. doi: 10.1371/journal.pbio.1002130
12. CDB. *Convention on Biological Diversity, Programme of Work on Invasive Alien Species*. Available online at: https://www.cbd.int/invasive (2016).
13. MEA (Millennium Ecosystem Assessment). *Ecosystems and Human Wellbeing: Biodiversity Synthesis.* Washington, DC: World Resources Institute (2005). p. 86.
14. CDB. Convention on biological diversity, interim guiding principles. In: *Conference of the Parties Decision V/8 Alien Species That Threaten Ecosystems, Habitats or Species*. Available online at: http://www.cbd.int/decision/cop/default.shtml?id=7150 (2000).
15. Katsanevakis S, Roy HE. Alien species related information systems and information management. *Manage Biol Invas.* (2015) **6**:115–17. doi: 10.3391/mbi.2015.6.2.01
16. Lucy FE, Roy H, Simpson A, Carlton JT, Hanson JM, Magellan K, et al. INVASIVESNET towards an international association for open knowledge on invasive alien species. *Manage Biol Invas.* (2016) **7**:131–9. doi: 10.3391/mbi.2016.7.2.01
17. Moustakas A, Voutsela A, Katsanevakis S. Sampling alien species inside and outside protected areas: does it matter? *SciTot Environ.* (2018) **625**:194–8. doi: 10.1016/j.scitotenv.2017.12.198

18. Evans MR, Moustakas A. A comparison between data requirements and availability for calibrating predictive ecological models for lowland UK woodlands: learning new tricks from old trees. *Ecol Evol.* (2016) **6**:4812–22. doi: 10.1002/ece3.2217
19. Evans MR, Benton TG, Grimm V, Lessells CM, O'Malley MA, Moustakas A, et al. Data availability and model complexity, generality, and utility: a reply to Lonergan. *Trends Ecol Evol.* (2014) **29:**302–3. doi: 10.1016/j.tree.2014.03.004
20. Moustakas A. Spatio-temporal data mining in ecological and veterinary epidemiology. *Stoch Environ Res Risk Assess.* (2017) **31**:829–34. doi: 10.1007/s00477-016-1374-8
21. Moustakas A, Evans MR. A big-data spatial, temporal and network analysis of bovine tuberculosis between wildlife (badgers) and cattle. *Stoch Environ Res Risk Assess.* (2017) **31**:315–28. doi: 10.1007/s00477-016-1311-x
22. Moustakas A, Evans M. Coupling models of cattle and farms with models of badgers for predicting the dynamics of bovine tuberculosis (TB). *Stoch Environ Res Risk Assess.* (2015) **29**:623–35. doi: 10.1007/s00477-014-1016-y
23. Cunze S, Koch LK, Kochmann J, Klimpel S. Aedes albopictus and Aedes japonicus - two invasive mosquito species with different temperature niches in Europe. *Parasit Vect.* (2016) **9**:573. doi: 10.1186/s13071-016-1853-2

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers
in Applied Mathematics and Statistics

PERSPECTIVE
published: 30 June 2017
doi: 10.3389/fams.2017.00013

Seven Recommendations to Make Your Invasive Alien Species Data More Useful

Quentin J. Groom[1]*, Tim Adriaens[2], Peter Desmet[2], Annie Simpson[3], Aaike De Wever[4], Ioannis Bazos[5], Ana Cristina Cardoso[6], Lucinda Charles[7], Anastasia Christopoulou[5], Anna Gazda[8], Harry Helmisaari[9], Donald Hobern[10], Melanie Josefsson[11], Frances Lucy[12], Dragana Marisavljevic[13], Tomasz Oszako[14], Jan Pergl[15], Olivera Petrovic-Obradovic[16], Céline Prévot[17], Hans P. Ravn[18], Gareth Richards[7], Alain Roques[19], Helen E. Roy[20], Marie-Anne A. Rozenberg[19], Riccardo Scalera[21], Elena Tricarico[22], Teodora Trichkova[23], Diemer Vercayie[24], Argyro Zenetos[25] and Sonia Vanderhoeven[26]

[1] Botanic Garden Meise, Bouchout Domain, Meise, Belgium, [2] Research Institute for Nature and Forest, Brussels, Belgium, [3] U.S. Geological Survey, Core Science Analytics, Synthesis, and Libraries Program, Reston, VA, United States, [4] Aquatic and Terrestrial Ecology, Operational Directorate Natural Environment, Royal Belgian Institute of Natural Sciences, Brussels, Belgium, [5] Department of Ecology and Systematics, Faculty of Biology, National and Kapodistrian University of Athens, Panepistimiopolis, Greece, [6] EASIN, European Commission Joint Research Centre, European Commission, Joint Research Centre, Ispra, Italy, [7] CABI, Nosworthy Way, Wallingford, Oxfordshire, United Kingdom, [8] Department of Forest Biodiversity, University of Agriculture in Krakow, Krakow, Poland, [9] Finnish Environment Institute, Helsinki, Finland, [10] Global Biodiversity Information Facility, Copenhagen, Denmark, [11] Swedish Environmental Protection Agency, Stockholm, Sweden, [12] CERIS, Institute of Technology, Sligo, Ireland, [13] Institute for Plant Protection and Environment, Belgrade, Serbia, [14] Forest Research Institute, Raszyn, Poland, [15] Department of Invasion Ecology, Institute of Botany, The Czech Academy of Sciences, Průhonice, Czechia, [16] Faculty of Agriculture, University of Belgrade, Belgrade, Serbia, [17] SPW-DEMNA, Département de l'Étude du Milieu Naturel et Agricole, Gembloux, Belgium, [18] Department of Geosciences and Natural Resource Management, Faculty of Science, University of Copenhagen, Frederiksberg, Denmark, [19] INRA, UR633, Zoologie Forestière, Orléans, France, [20] Centre for Ecology and Hydrology, Wallingford, United Kingdom, [21] IUCN SSC Invasive Species Specialist Group, Rome, Italy, [22] Università degli Studi di Firenze, Firenze, Italy, [23] Institute of Biodiversity and Ecosystem Research, Bulgarian Academy of Sciences, Sofia, Bulgaria, [24] Natuurpunt, Mechelen, Belgium, [25] Institute of Marine Biological Resources and Inland Waters, HCMR, Anavyssos, Greece, [26] Belgian Biodiversity Platform, Walloon Research Department for Nature and Agricultural Areas, Service Public de Wallonie, Gembloux, Belgium

Edited by:
Aristides Moustakas,
Queen Mary University of London,
United Kingdom

Reviewed by:
Sarah Supp,
University of Wisconsin-Madison,
United States
Davide Francesco Tagliapietra,
ISMAR-Marine Sciences Institute in
Venice (CNR), Italy

***Correspondence:**
Quentin J. Groom
quentin.groom@plantentuinmeise.be

Specialty section:
This article was submitted to
Environmental Informatics,
a section of the journal
Frontiers in Applied Mathematics and
Statistics

Received: 12 March 2017
Accepted: 15 June 2017
Published: 30 June 2017

Citation:
Groom QJ, Adriaens T, Desmet P, Simpson A, De Wever A, Bazos I, Cardoso AC, Charles L, Christopoulou A, Gazda A, Helmisaari H, Hobern D, Josefsson M, Lucy F, Marisavljevic D, Oszako T, Pergl J, Petrovic-Obradovic O, Prévot C, Ravn HP, Richards G, Roques A, Roy HE, Rozenberg M-AA, Scalera R, Tricarico E, Trichkova T, Vercayie D, Zenetos A and Vanderhoeven S (2017) Seven Recommendations to Make Your Invasive Alien Species Data More Useful. Front. Appl. Math. Stat. 3:13. doi: 10.3389/fams.2017.00013

Science-based strategies to tackle biological invasions depend on recent, accurate, well-documented, standardized and openly accessible information on alien species. Currently and historically, biodiversity data are scattered in numerous disconnected data silos that lack interoperability. The situation is no different for alien species data, and this obstructs efficient retrieval, combination, and use of these kinds of information for research and policy-making. Standardization and interoperability are particularly important as many alien species related research and policy activities require pooling data. We describe seven ways that data on alien species can be made more accessible and useful, based on the results of a European Cooperation in Science and Technology (COST) workshop: (1) Create data management plans; (2) Increase interoperability of information sources; (3) Document data through metadata; (4) Format data using existing standards; (5) Adopt controlled vocabularies; (6) Increase data availability; and (7) Ensure long-term data preservation. We identify four properties specific and integral to alien species data (species status, introduction pathway, degree of establishment, and impact mechanism) that are either missing from existing data standards or lack a recommended controlled vocabulary. Improved access to accurate, real-time and historical data will repay the long-term investment in data management infrastructure, by providing more accurate, timely and realistic assessments and analyses. If we improve core biodiversity data standards by developing their relevance to alien species, it will allow the automation of common activities regarding data processing in support of environmental policy. Furthermore, we call for considerable effort to maintain, update, standardize, archive, and aggregate datasets, to ensure proper valorization of alien species data and information before they become obsolete or lost.

Keywords: checklists, data interoperability, data management plan, introduced species, non-indigenous, non-native, pest species, standards

1. INTRODUCTION

Sound decision-making to minimize the risk associated with the introduction of alien species requires accurate and up-to-date data and the knowledge derived from them. These data feed into a wide range of processes to tackle problematic invasive alien species and are needed to develop an appropriate, evidence-based response (**Table 1**). Horizon scanning (the systematic examination of future potential threats and opportunities, leading to their prioritization), risk assessment, risk management, early detection and rapid response all depend on accurate and accessible data [1–4]. So, although alien species data are little different from data on other species, the demands we place on these data are considerable and specific.

Current invasive alien species policies depend on the availability and quality of data. For example, the EU regulation no. 1143/2014 on Invasive Alien Species [5], requires member states to report on the status of invasive alien species of Union concern and their progress in managing them, likewise similar regulations exist in other countries, such as the USA [6]. Responsible authorities need access to timely and validated data and they need to report this in a standardized way, so it can be collated nationally and internationally. Within the EU, the European Alien Species Information Network (EASIN) [7, 8] has been developed to this end, including a mechanism for quality assurance, safeguarding and improvement [9].

Mitigating and preventing biological invasions present particular challenges with regard to the quality, relevance and scope of data sources and infrastructure [10]. The numerous origins of the data and broad taxonomic scope, combined with the global geographic extent and input from diverse disciplines make proper handling of alien species data difficult, but also necessary. With this perspective, we gathered database managers, data users, data generators and biodiversity informatics specialists to outline how alien species data can be made more useful, taking into account the peculiarities and applications of such data. This resulted in seven recommendations, which, if followed, would improve the use of alien species data for research, policy and management purposes. Some of these recommendations are not unique to alien species data, but their impact would be particularly significant in this discipline.

TABLE 1 | Data/information categories and their invasive alien species-related evidence-based processes.

Alien species checklists	Horizon scanning (e.g., [2]) Selection of species for risk assessment (e.g., [46]) Analysis of pathways of introduction and spread (e.g., [43, 61, 62]) Pathway regulation Feeding indicators for policy evaluation (e.g., [63–65])
Occurrence data of alien and native species	Species distribution models (e.g., [66, 67]) Niche and occupancy modelingRisk modeling and risk mapping (e.g., [68]) Impact assessment Risk assessment Climate matching Impact research (e.g., [69, 70]) Early warning and rapid response programs (e.g., [35, 71])
Climate data	Niche and occupancy modeling Climate matching (e.g., [72]) Risk assessment
Genetic data	Species identification (e.g., [71, 73]) Early detection through e-DNA (e.g., [32])
Data on management actions	Risk management (e.g., [62]) Evaluation of effectiveness of control actions Cost-benefit analysis of control actions (e.g., [74, 75]) Assessment of non-target effects of control actions

2. APPROACH

Correspondence A workshop titled *Data for invasive species research, policy making and management* was organized in Brussels in 2016 with representatives from sixteen European countries and the United States. The attendees were from the European Alien Challenge COST[1] Action, from important institutions and projects related to alien species data such as

---

[1] http://www.cost.eu/

the European Alien Species Information Network (EASIN), Delivering Alien Invasive Species Inventories for Europe (DAISIE), Global Biodiversity Information Facility (GBIF), Global Invasive Species Information Network (GISIN), Centre for Agriculture and Biosciences International Invasive Species Compendium (CABI–ISC), and the Biodiversity Information Standards organization (formerly known as the Taxonomic Databases Working Group and referred to by the acronym TDWG). Effort was made to balance participant representation in terms of gender, country of origin within Europe and taxonomic and habitat interests (terrestrial, freshwater and marine).

The workshop consisted of talks and participatory exercises on four main invasive alien species themes: risk assessment, horizon scanning, management and monitoring. For each of these themes, participants reflected on the data needs and requirements (**Table 1**), the data sources they commonly use, and the existing data standards. Materials from the workshop have been deposited in an open repository [11]. Conclusions reported by breakout groups were refined and supplemented in facilitated plenary discussion. Particular attention was paid to the perspectives of both the data publishers and data users.

During the workshop a number of opportunities for facilitating proper use and valorization of alien species data was identified and these resulted in the recommendations presented below and summarized in **Table 2**.

3. CREATE DATA MANAGEMENT PLANS

A DMP describes how the information generated by a project will be handled both during and after it is generated. These plans define responsibilities; aim to avoid data loss and incompatibility by indicating how data will be preserved and formatted; stipulate what metadata are required to understand the data; and consider data sharing options, including licensing [12].

Such plans are a means to improve data management and are now commonly required by funding agencies. The US National Science Foundation has required them since 2010 [13] and in 2013 the European Commission launched a pilot on open research data requiring a DMP in the first 6 months of the project [14]. The DMP approach also encourages journals to change their policies toward the archiving of data, though it is taking time for the whole scientific community to embrace these changes [15, 16]. Typical minimum sections of a DMP are: (i) What type of data and metadata are expected? (ii) Which standards are used for alien species data? (iii) How should data be shared? (iv) How should data be permanently preserved? Researchers new to writing a DMP should refer to their institutional and funding agency guidelines if any, and, with respect to invasive species data, recommendations for ecologists [6, 17].

Strictly speaking, each recommended action could be implemented without the need to compile a DMP. However, preparing and agreeing upon a DMP ensures a holistic approach to data management and increases its openness and accountability, while also answering the needs from funding agencies and institutional data policies [12], so we recommend their use.

4. DOCUMENT THROUGH METADATA

Good metadata provide information on provenance, scope, methods, limitations, data formats and units to facilitate correct data use, as well as license and contact information. USGS' Data Management Web site[2] lists multiple tools and best practices for metadata creation. Several metadata standards for biodiversity data are available: such as Ecological Markup Language (EML [18]) adopted by GBIF [19]; the INSPIRE directive framework (Infrastructure for Spatial Information in Europe)[3], which describes geospatial data and the Data Catalog Vocabulary (DCAT)[4], to describe datasets. We have not identified any specific metadata standards for alien species data and recommend the use of the metadata standards above, for which tools and services are already available [20]. An example of a tool for metadata standardization is the desktop application Morpho[5], which guides users through the creation of EML [21]. Morpho can interface with a MetaCat registry to provide a searchable catalog of ecological datasets. This technology is used by both the DataONE repository[6] and the European Biological Observations Network (EU BON) [22]. Creating metadata may seem secondary to primary data curation, but metadata are essential to ensure the data

TABLE 2 | Seven recommendations for improving the usefulness of alien species data.

1	Create and implement data management plans to define the alien species data life-cycle, good data quality and metadata, standardization, data sharing options, and long-term data preservation.
2	Describe alien species data through metadata, so users can understand their scope and limitations, and use metadata standards (EML, INSPIRE) to facilitate metadata exchange.
3	Improve interoperability and sustainability of existing and new alien species information sources by exposing the data they contain through standard exchange formats.
4	Format data using existing standards (Darwin Core, GISIN) and engage in their development through TDWG.
5	Adopt controlled vocabularies to further increase interoperability of data and engage with TDWG to make these compatible with existing standards.
6	Increase data availability by making alien species data openly accessible as soon as possible after collection.
7	Ensure long-term preservation of alien species data by archiving these in existing data repositories (GBIF, Zenodo).

---

[2] https://www2.usgs.gov/datamanagement/describe/metadata.php
[3] http://inspire.ec.europa.eu
[4] https://www.w3.org/TR/vocab-dcat/
[5] https://knb.ecoinformatics.org/#tools/morpho
[6] http://dataone.org

can be discovered and used in the long term [23]. In the context of alien species data, improved access to metadata could enhance the speed with which data are found and mobilized.

5. IMPROVED INTEROPERABILITY OF INFORMATION SOURCES

Information on alien species is scattered among a multitude of sources, including databases; peer-reviewed and gray literature; unpublished research projects and institutional datasets [8, 24]. Important international sources of these data include the 2000 Global Invasive Species Database (GISD) of the IUCN/SSC Invasive Species Specialist Group (ISSG) [25]; the 2004 Global Invasive Species Information Network (GISIN); and the Global Invasive Alien Species Information Partnership (GIASIP), as well as global information providers such as the CABI Invasive Species Compendium (ISC) and the Global Register of Introduced and Invasive Species (GRIIS). Any new initiative to collate data needs to consider its role and define its niche within a complex environment of global, continental, national and regional data repositories [7, 26].

Almost any effort to compile and harmonize data from these sources is impeded by differences in field names, definitions, and taxonomy, as well as access and license restrictions [3, 27]. The use of common standards for all these aspects can improve the interoperability of these data sources: their data can be more efficiently exchanged, combined, compared, and presented. In addition, data processing should ideally be performed in a repeatable way, to increase the efficiency of activities such as horizon scanning and risk assessment. For invasion policies to be proactive, these activities should be repeated at regular intervals [2].

Online alien species catalogs and invasive alien species information systems are difficult to keep up-to-date [28, 29], yet they provide a wide variety of valuable information. Funding for these initiatives has been sporadic at best [28] and is often time-limited [29]. Thus relevant information stored and managed within such initiatives become quickly out-dated, and efforts to keep them updated are often suddenly discontinued. This tends to spread errors to other systems that are populated with data from such sources, particularly if provenance is poorly tracked. As such, the current process restricts alien species data exchange, aggregation, interoperability and even rescue. Technological advances have boosted the number of initiatives [30], but also increased the data's volume and complexity [23, 31–33]. A holistic approach to complex biological questions requires more from data than a traditional reductionist approach, as demonstrated by the success of the Gene Ontology [34]. Yet this poses additional challenges of ensuring data quality, data curation, interoperability and future-proofing against obsolete technology and increasing data volumes [35]. Technological change promises many improvements in data collection, with systems such as smartphones equipped with built-in GPS, image capture, external sensors, and automated and expert validation [31]. Also, advances in species detection through environmental DNA, such as those of Dejean et al. [32], need support to be included within alien species initiatives.

We recommend that alien species databases work together to follow common standards and that these standards are further developed for emerging data streams.

6. FORMAT DATA USING EXISTING STANDARDS

Within the scope of a single dataset, data only need to be formatted consistently to be usable. However, to combine datasets for broad-scale analysis, a community-defined exchange format or standard is required to allow data interoperability. Among the qualities of a "good standard" are that it be readable (by both humans and machines), simple, learnable and efficient [36].

The alien species research community is not universally aware of biodiversity informatics standards, where they come from and how they can be extended. Standards for the exchange of biodiversity data, including alien species data, are developed, discussed and promoted by the Biodiversity Information Standards organization, TDWG [37]. This organization is the guardian of Darwin Core, the most widely adopted standard to exchange biodiversity information related to species [38]. By following these standards, data managers can avoid duplication of effort and mistakes. Furthermore, the organization can give advice and support for updating existing standards and proposing new ones. It is recommended that the invasive alien species community continue to engage in TDWG, both to adopt standards for common terms and to establish standards specific to invasion biology.

7. ADOPT CONTROLLED VOCABULARIES FOR FOUR ALIEN SPECIES PROPERTIES

In addition to a standard format to exchange data, specialist communities often also require further control on the values of terms to increase interoperability. This can be achieved by adopting controlled vocabularies. This not only means that data can be merged, but also contributes to the normative definition of a term.

Four alien species properties were identified that are either missing from Darwin Core or lacking a reference to a recommended controlled vocabulary. These are introduction pathway, degree of establishment, impact mechanism, and species status. For each of these, vocabularies exist outside Darwin Core, yet these currently exist as frameworks and require further work to be developed into standards.

For pathway terminology, the need for a consistent classification, hierarchy, and terminology has long been recognized [39–41]. Meanwhile, a standardized hierarchical pathway classification was adopted by parties to the Convention of Biological Diversity [42] and is being applied to existing databases [9, 43].

A framework for the degree of establishment has been presented by Blackburn et al. [44]. This hierarchical classification provides a terminology for populations at different points in the invasion process (casual/introduced, alien, naturalized/established, invasive) and allows expression of the range of establishments from those organisms only kept in cultivation or captivity through to full naturalization and invasiveness.

For alien species impact, a classification of categories based on the magnitude of environmental impacts was developed by Blackburn et al. [45], and has been adopted by the IUCN in 2016. However, for impacts other than environmental, such as socio-economic, plant health, human health and animal health, no comprehensive overview is available, but several protocols have been developed for risk assessment that can provide inspiration for classifications (see [46] for an overview).

Standards from the trade and agriculture sectors can be useful in describing species status, for example, the International Plant Protection Conventions International Standard for Phytosanitary Measures: specifically, IPSM 8[7], Determining pest status in an area; and IPSM 6[8], Guidelines for surveillance. We recommend these controlled vocabularies are expressed in a machine-readable format and are referenced from the appropriate terms in Darwin Core. This is in line with the recommendations of the GBIF Task Group on Data Fitness for use in Research into Invasive Alien Species [33].

Additionally, controlled vocabularies might prove helpful in the dissemination of information on species management [47]. Good examples are the Global Eradication and Response Database [48] and the Database of Island Invasive Species Eradications [49]. The documentation of management actions in the field and the storage of these data are key to performing cost-benefit analyses of management measures.

8. INCREASE DATA AVAILABILITY

Much has already been written about the methods and needs for open data publication [3, 17, 50]. Beyond the good intentions, Invasivesnet is a developing global association for open knowledge and open data on alien species [51]. This association will facilitate greater understanding, communication, and improved management of biological invasions globally, by developing a sustainable network of networks for effective knowledge exchange. The association fosters tool development and cyberinfrastructure for the collection, management and dissemination of data and information on alien species from a range of sources (e.g., research, citizen science). The key point is that data should be shared and standardized to ensure interoperability [52]. In the case of species observation data a straightforward solution is to publish through a repository such as GBIF or the Ocean Biogeographic Information System (OBIS), as it ensures adherence to a minimum of common standards.

There can be little doubt that data sharing using community standards and adequate metadata are of benefit to research and society in general [53]. Yet motivating good data management is not easy when practitioners are not rewarded by their institutions. However, this is changing [54, 55], particular with the support of aspirational statements such as the Berlin[9] and Bouchout[10] declarations, which show the willingness of some institutions and individuals to change. Also, there are now policy initiatives in place, such as the EU INSPIRE directive[11] or the United States Administration's Open Data Policy[12,13], to mandate harmonization of spatial data.

9. ENSURE LONG-TERM DATA PRESERVATION

Under ideal circumstances databases would have funding for maintenance and updating for as long as they are useful, however, this is unrealistic. Furthermore, the end of a project is the wrong time to consider the long-term persistence of data [29, 56]. Data actively being curated are often best maintained close to their source, however, longevity can be built-in to procedures by periodically depositing data in an open repository, not just on a personal or university website. Hence, data are protected from catastrophic events, human attrition, and the slow degradation of obsolescent hardware, which is the fate of much data [57]. If a publication is based upon a specific dataset it is good practice to deposit that precise version in a repository.

Not all repositories are the same, for example the Dryad[14] and Zenodo[15] repositories are general-purpose repositories able to accept data in *ad hoc* formats, not necessarily formatted to community standards. They provide flexibility, however, repositories dedicated to one data type provide much greater opportunities for integration due to their enforcement of standards. Examples of such repositories are GBIF and GenBank [58]. Repositories also differ in their ability to embargo the release of data and in the licensing options. We recommend that considerable *a priori* thought goes into data preservation and the choice of repository.

10. CONCLUSION

Many alien species databases have emerged either before or without knowledge of existing standards for database management in biodiversity informatics. Furthermore, existing standards do not adequately cover all the needs of the research domain. Not all ecologists have strong information technology skills, nor are experts in technology-mediated collaboration, shared instrumentation or standardized data collection [59]. In the rapidly changing information technology landscape, ecologists and conservationists cannot be expected to keep up with developments in software and data standards. This should encourage data managers, wherever possible, to simplify the tools

---

[7] https://www.ippc.int/en/publications/612/
[8] https://www.ippc.int/en/publications/615/
[9] https://openaccess.mpg.de/Berlin-Declaration
[10] http://bouchoutdeclaration.org/
[11] http://inspire.ec.europa.eu
[12] https://www.fgdc.gov/standards
[13] https://project-open-data.cio.gov/IDC/
[14] http://datadryad.org/
[15] https://zenodo.org/

created for ecological practitioners. This becomes more pressing as new technologies are used to provide data on alien species.

Many data management issues are common to all biodiversity data, yet species native range, introduction pathway, degree of establishment and impact mechanism are specific to alien species. Additionally, the need for fast dissemination of information and data is typical to alien species, in particular early detection and rapid response programs. Proactive responses to biological invasions require repeatable workflows for horizon scanning and risk assessment [60]. Adoption of standards and controlled vocabularies for this information can boost the usefulness for alien species research, policy-making and policy evaluation. There is a need for the acceptance of common data standards that take into consideration the needs of both data collectors and diverse data users, from the science community to the end user.

Work is required with the research and education communities and the standards authorities to ensure that suggested standards are shepherded through acceptance and implementation and that these standards are introduced early within the education of young scientists and promoted among those in the biodiversity community, so that they are adopted widely. Improving core biodiversity standards for their content and usefulness for alien species data will allow the automation of common activities needed to tackle biological invasions. We call for considerable effort toward maintaining, updating, standardizing, and archiving or incorporating current data sets, to ensure proper valorization of alien species data and resulting information before they become obsolete or lost.

AUTHOR CONTRIBUTIONS

QG, ACC, JP, SV, and TA wrote the original briefing note, which outlined the idea of a workshop on biodiversity data interoperability for invasive species. SV, QG, TA, PD, AD were the local organizers of the Workshop and prepared the initial draft of the paper. HR is Chair of the COST Action and has supported and attended the workshop. AS participated in the workshop, contributed to the writing of the paper, and arranged for the initial peer review of the manuscript through the U.S. Geological Survey. All other authors contributed to the writing of the paper and attended the workshop.

FUNDING

This article is based upon work from COST Action TD1209 ALIEN Challenge, supported by COST (European Cooperation in Science and Technology) www.cost.eu. JP was partly supported by the long-term research development project no. RVO 67985939.

ACKNOWLEDGMENTS

We also thank the Belgian Biodiversity Platform, funded by the Belgian Science Policy Office, for the use of their facilities, and the U.S. Geological Survey, for previous peer review. Any use of trade, product, or firm names in this article is for descriptive purposes only and does not imply endorsement by the U.S. Government.

REFERENCES

1. McGeoch MA, Spear D, Kleynhans EJ, Marais E. Uncertainty in invasive alien species listing. *Ecol Appl.* (2012) **22**:959–71. doi: 10.1890/11-1252.1
2. Roy HE, Peyton J, Aldridge DC, Bantock T, Blackburn TM, Britton R, et al. Horizon scanning for invasive alien species with the potential to threaten biodiversity in Great Britain. *Global Change Biol.* (2014) **20**:3859–71. doi: 10.1111/gcb.12603
3. Groom QJ, Desmet P, Vanderhoeven S, Adriaens T. The importance of open data for invasive alien species research, policy and management. *Manage Biol Invasions* (2015) **6**:119–25. doi: 10.3391/mbi.2015.6.2.02
4. Zenetos A, Çinar ME, Crocetta F, Golani D, Rosso A, Servello G, et al. Uncertainties and validation of alien species catalogues: the Mediterranean as an example. *Estuarine Coast Shelf Sci.* (2017) **191**:171–87. doi: 10.1016/j.ecss.2017.03.031
5. European Union. Regulation (EU) No. 1143/2014 of the European Parliament and of the Council on the prevention and management of the introduction and spread of invasive alien species. *Official J Eur Union* (2014) **317**:35–55.
6. Executive Office of the President. Safeguarding the nation from the impacts of invasive species. *Federal Register* (2016) **81**:88609–14.
7. Katsanevakis S, Bogucarskis K, Gatto F, Vandekerkhove J, Deriu I, Cardoso AC, et al. Building the European Alien Species Information Network (EASIN): a novel approach for the exploration of distributed alien species data. *Bioinvasions Records* (2012) **1**:235–45. doi: 10.3391/bir.2012.1.4.01
8. Katsanevakis S, Deriu I, D'Amico F, Nunes AL, Pelaez Sanchez S, Crocetta F, et al. European alien species information network (EASIN): supporting European policies and scientific research. *Manage Biol Invasions* (2015) **6**:147–57. doi: 10.3391/mbi.2015.6.2.05
9. Tsiamis K, Cardoso AC, Gervasini E. The European Alien Species Information Network on the Convention on Biological Diversity pathways categorization. *NeoBiota* (2017) **2**:21–29. doi: 10.3897/neobiota.32.9429
10. Latombe G, Pyšek P, Jeschke JM, Blackburn TM, Bacher S, Capinha C, et al. A vision for global monitoring of biological invasions. *Biol Conserv*. (2016). doi: 10.1016/j.biocon.2016.06.013. [Epub ahead of print].
11. Groom Q, Adriaens T, Desmet P, Wever AD, Vanderhoeven S. *Resources from the Workshop "Data for Invasive Species Research, Policy Making and Management"*. doi: 10.5281/zenodo.261743
12. McLure M, Level AV, Cranston CL, Oehlerts B, Culbertson M. Data curation: a study of researcher practices and needs. *Libraries Acad.* (2014) **14**:139–64. doi: 10.1353/pla.2014.0009
13. National Science Foundation. *Press Release 10-077 Scientists Seeking NSF Funding Will Soon be Required to Submit Data Management Plans.* (2010). Available online at: http://www.nsf.gov/news/news_summ.jsp?cntn_id=116928 (Accessed 2 October, 2010).
14. European Commission. *EC Commission Recommendation of 17.7.2012 on Access to and Preservation of Scientific Information*. (2012). Available online at: http://ec.europa.eu/research/science-society/document_library/pdf_06/recommendation-access-and-preservation-scientific-information_en.pdf (Accessed 4 November, 2016).
15. Roche DG, Kruuk LEB, Lanfear R, Binning SA. Public data archiving in ecology and evolution: how well are we doing? *PLoS Biol.* (2015) **13**:e1002295. doi: 10.1371/journal.pbio.1002295
16. Vision TJ. Open data and the social contract of scientific publishing. *Bioscience* (2010) **60**:330. doi: 10.1525/bio.2010.60.5.2
17. White EP, Baldridge E, Brym ZT, Locey KJ, McGlinn DJ, Supp SR. Nine simple ways to make it easier to (re) use your data. *Ideas Ecol Evol.* (2013) **6**:1–10. doi: 10.4033/iee.2013.6b.6.f
18. Fegraus EH, Andelman S, Jones MB, Schildhauer M. Maximizing the value of ecological data with structured metadata: an introduction to ecological metadata language (EML) and principles for metadata creation. *Bull Ecol Soc Am*. (2005) **86**:158–68. doi: 10.1890/0012-9623(2005)86[158:MTVOED]2.0.CO;2

19. Robertson T, Dring M, Guralnick R, Bloom D, Wieczorek J, Braak K, et al. The GBIF integrated publishing toolkit: facilitating the efficient publishing of biodiversity data on the internet. *PLoS ONE* (2014) **9**:e102623. doi: 10.1371/journal.pone.0102623
20. Oggioni A, Carrara P, Kliment T, Peterseil J, Schentz H. Monitoring of environmental status through long term series: data management system in the EnvEurope project. In: *EnviroInfo*. Aachen: Shaker Verlag (2012). p. 293–301.
21. Higgins D, Berkley C, Jones MB. Managing heterogeneous ecological data using Morpho. In: *Proceedings of the 14th International Conference on Scientific and Statistical Database Management*. IEEE (2002). p. 69–76.
22. Smirnova L, Mergen P, Groom Q, Wever AD, Penev L, Stoev P, et al. Data sharing tools adopted by the European Biodiversity Observation Network Project. *Res Ideas Outcomes* (2016) **2**:e9390. doi: 10.3897/rio.2.e9390
23. Michener WK. Meta-information concepts for ecological data management. *Ecol Inform*. (2006) **1**:3–7. doi: 10.1016/j.ecoinf.2005.08.004
24. Roy HE, Preston CD, Harrower CA, Rorke SL, Noble D, Sewell J, et al. GB Non-native Species Information Portal: documenting the arrival of non-native species in Britain. *Biol Invasions* (2014) **16**:2495–505. doi: 10.1007/s10530-014-0687-0
25. Pagad S, Genovesi P, Carnevali L, Scalera R, Clout M. IUCN SSC Invasive Species Specialist Group: invasive alien species information management supporting practitioners, policy makers and decision takers. *Manage Biol Invasions* (2015) **6**:127–35. doi: 10.3391/mbi.2015.6.2.03
26. Katsanevakis S, Genovesi P, Gaiji S, Hvid HN, Roy H, Nunes AL, et al. Implementing the European policies for alien species: networking, science, and partnership in a complex environment. *Manage Biol Invasions* (2013) **4**:3–6. doi: 10.3391/mbi.2013.4.1.02
27. Joppa LN, O'Connor B, Visconti P, Smith C, Geldmann J, Hoffmann M, et al. Filling in biodiversity threat gaps. *Science* (2016) **352**:416–8. doi: 10.1126/science.aaf3565
28. Jarnevich CS, Simpson A, Graham JJ, Newman GJ, Bargeron CT. Running a network on a shoestring: the Global Invasive Species Information Network. *Manage Biol Invasions* (2015) **6**:137–46. doi: 10.3391/mbi.2015.6.2.04
29. Olenin S, Narius A, Minchin D, David M, Galil B, Gollasch S, et al. Making non-indigenous species information systems practical for management and useful for research: an aquatic perspective. *Biol Conserv*. (2014) **173**:98–107. doi: 10.1016/j.biocon.2013.07.040
30. Roy H, Pocock M, Preston C, Roy D, Savage J, Tweddle J, et al. *Understanding Citizen Science and Environmental Monitoring: Final Report on Behalf of UK Environmental Observation Framework*. NERC/Centre for Ecology and Hydrology (2012).
31. August T, Harvey M, Lightfoot P, Kilbey D, Papadopoulos T, Jepson P. Emerging technologies for biological recording. *Biol J Linnean Soc*. (2015) **115**:731–49. doi: 10.1111/bij.12534
32. Dejean T, Valentini A, Miquel C, Taberlet P, Bellemain E, Miaud C. Improved detection of an alien invasive species through environmental DNA barcoding: the example of the American bullfrog Lithobates catesbeianus. *J Appl Ecol*. (2012) **49**:953–9. doi: 10.1111/j.1365-2664.2012.02171.x
33. McGeoch MA, Groom QJ, Pagad S, Petrosyan V, Ruiz G, Wilson J. *Data Fitness for Use in Research on Alien and Invasive Species*. Copenhagen: GBIF Secretariat (2016). Available online at: http://www.gbif.org/resource/82958
34. Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, et al. Gene ontology: tool for the unification of biology. *Nat Genet*. (2000) **25**:25–9. doi: 10.1038/75556
35. Simpson A, Jarnevich C, Madsen J, Westbrooks R, Fournier C, Mehrhoff L, et al. Invasive species information networks: collaboration at multiple scales for prevention, early detection, and rapid response to invasive alien species. *Biodiversity* (2009) **10**:5–13. doi: 10.1080/14888386.2009.9712839
36. Bos B. *What Is a Good Standard? An Essay on W3C's Design Principles.* (2003). Available online at: https://www.w3.org/People/Bos/DesignGuide/introduction
37. Berendsohn W, Güntsch A, Hoffmann N, Kohlbecker A, Luther K, Müller A. Biodiversity information platforms: from standards to interoperability. *ZooKeys* (2011) **150**:71–87. doi: 10.3897/zookeys.150.2166
38. Wieczorek J, Bloom D, Guralnick R, Blum S, Döring M, Giovanni R, et al. Darwin core: an evolving community-developed biodiversity data standard. *PLoS ONE* (2012) **7**:e29715. doi: 10.1371/journal.pone.0029715
39. Hulme PE, Bacher S, Kenis M, Klotz S, Kühn I, Minchin D, et al. Grasping at the routes of biological invasions: a framework for integrating pathways into policy. *J Appl Ecol*. (2008) **45**:403–14. doi: 10.1111/j.1365-2664.2007.01442.x
40. Essl F, Bacher S, Blackburn TM, Booy O, Brundu G, Brunel S, et al. Crossing frontiers in tackling pathways of biological invasions. *Bioscience* (2015) **65**:769. doi: 10.1093/biosci/biv082
41. Pagad S, Hayes K, Katsanevakis S, Costello MJ. *World Register of Introduced Marine Species* (WRIMS) (2015). Available online at: http://www.marinespecies.org/introduced
42. Convention on Biological Diversity. *Pathways of Introduction of Invasive Species, Their Prioritization, and Management* (2014). Available online at: http://www.cbd.int/doc/meetings/sbstta/sbstta-18/official/sbstta-18-09-add1-en.pdf
43. Saul WC, Roy HE, Booy O, Carnevali L, Chen HJ, Genovesi P, et al. Assessing patterns in introduction pathways of alien species by linking major invasion data bases. *J Appl Ecol*. (2017) **54**:657–69. doi: 10.1111/1365-2664.12819
44. Blackburn TM, Pyšek P, Bacher S, Carlton JT, Duncan RP, Jarošík V, et al. A proposed unified framework for biological invasions. *Trends Ecol Evol*. (2011) **26**:333–9. doi: 10.1016/j.tree.2011.03.023
45. Blackburn TM, Essl F, Evans T, Hulme PE, Jeschke JM, Kühn I, et al. A unified classification of alien species based on the magnitude of their environmental impacts. *PLoS Biol*. (2014) 12:e1001850. doi: 10.1371/journal.pbio.1001850
46. McGeoch MA, Genovesi P, Bellingham PJ, Costello MJ, McGrannachan C, Sheppard A. Prioritizing species, pathways, and sites to achieve conservation targets for biological invasion. *Biol Invasions* (2016) **18**:299–314. doi: 10.1007/s10530-015-1013-1
47. Hulme PE, Pyšek P, Nentwig W, Vilà M. Will threat of biological invasions unite the European Union? *Science* (2009) **324**:40–1. doi: 10.1126/science.1171111
48. Kean J, Tobin P, Lee D, Smith G, Stringer L, Vargas RF, et al. *Global Eradication and Response Database* (2016). Available online at: http://b3.net.nz/gerda
49. Island Conservation, Coastal Conservation Action Laboratory UCSC, IUCN SSC Invasive Species Specialist Group, University of Auckland and Landcare Research New Zealand. DIISE. *The Database of Island Invasive Species Eradications* (2015). Available online at: http://diise.islandconservation.org
50. Wilkinson MD, Dumontier M, Aalbersberg IJ, Appleton G, Axton M, Baak A, et al. The FAIR guiding principles for scientific data management and stewardship. *Sci Data* (2016) **3**:160018. doi: 10.1038/sdata.2016.18
51. Lucy FE, Roy H, Simpson A, Carlton JT, Hanson JM, Magellan K, et al. INVASIVESNET towards an international association for open knowledge on invasive alien species. *Manage Biol Invasions* (2016) **7**:131–9. doi: 10.3391/mbi.2016.7.2.01
52. Faith D, Collen B, Ariño A, Koleff PKP, Guinotte J, Kerr J, et al. Bridging the biodiversity data aps: recommendations to meet users' data needs. *Biodiversity Inform.* (2013) 8:41–58. doi: 10.17161/bi.v8i2.4126
53. Borgman CL. The conundrum of sharing research data. *J Am Soc Inform Sci Technol*. (2012) **63**:1059–78. doi: 10.1002/asi.22634
54. Donovan C. Introduction: future pathways for science policy and research assessment: metrics vs peer review, quality vs impact. *Sci Public Policy* (2007) **34**:538. doi: 10.3152/030234207X256529
55. Taylor J. The assessment of research quality in UK universities: peer review or metrics? *Brit J Manage*. (2011) **22**:202–17. doi: 10.1111/j.1467-8551.2010.00722.x
56. Maday C, Moysan M. Records management for scientific data. *Arch Manuscripts* (2014) **42**:190–2. doi: 10.1080/01576895.2014.911686
57. Gibney E, Van Noorden R. Scientists losing data at a rapid rate. *Nature* (2013) **504**. doi: 10.1038/nature.2013.14416
58. Benson DA, Cavanaugh M, Clark K, Karsch-Mizrachi I, Lipman DJ, Ostell J, et al. GenBank. *Nucleic Acids Res*. (2013) **41**:D36. doi: 10.1093/nar/gks1195
59. Ribes D, Finholt TA. The long now of technology infrastructure: articulating tensions in development. *J Assoc Inform Syst*. (2009) **10**:375.
60. Vanderhoeven S, Adriaens T, Desmet P, Strubbe D, Backeljau T, Barbier Y, et al. Tracking Invasive Alien Species (TrIAS): building a data-driven framework to inform policy. *Res Ideas Outcomes* (2017) **3**:e13414. doi: 10.3897/rio.3.e13414
61. NOBANIS. *Invasive Alien Species: Pathway Analysis and Horizon Scanning for Countries in Northern Europe*. Copenhagen: Nordic Council of Ministers (2015).

62. Andow DA. Pathways-based risk assessment of exotic species invasions. In: Carlton J, editor. *Invasive Species: Vectors and Management Strategies.* Washington, DC: Island Press (2003). p. 439–55.
63. Rabitsch W, Essl F, Genovesi P, Scalera R. *Invasive Alien Species Indicators in Europe-a Review of Streamlining European Biodiversity (SEBI) Indicator 10*. EEA-Technical Report, 15/2012 (2012).
64. Armon RH, Zenetos A. Invasive alien species and their indicators. In: Armon RH, Hänninen O, editors. *Environmental Indicators*. Dordrecht: Springer Netherlands (2015). p. 147–73. doi: 10.1007/978-94-017-9499-2_10
65. Rabitsch W, Genovesi P, Scalera R, Biała K, Josefsson M, Essl F. Developing and testing alien species indicators for Europe. *J Nat Conserv*. (2016) **29**:89–96. doi: 10.1016/j.jnc.2015.12.001
66. Ficetola GF, Thuiller W, Miaud C. Prediction and validation of the potential global distribution of a problematic alien invasive species—the American bullfrog. *Divers Distribut*. (2007) **13**:476–85. doi: 10.1111/j.1472-4642.2007.00377.x
67. Strubbe D, Matthysen E, Graham CH. Assessing the potential impact of invasive ring-necked parakeets Psittacula krameri on native nuthatches Sitta europeae in Belgium. *J Appl Ecol*. (2010) **47**:549–57. doi: 10.1111/j.1365-2664.2010.01808.x
68. Jiménez-Valverde A, Peterson AT, Soberón J, Overton J, Aragón P, Lobo JM. Use of niche models in invasive species risk assessments. *Biol Invasions* (2011) **13**:2785–97. doi: 10.1007/s10530-011-9963-4
69. Elzinga CL, Salzer DW, Willoughby JW, Gibbs JP. *Monitoring Plant and Animal Populations*. Malden, MA: Blackwell Science (2001).
70. Marsh DM, Trenham PC. Current trends in plant and animal population monitoring. *Conserv Biol*. (2008) **22:**647–55. doi: 10.1111/j.1523-1739.2008.00927.x
71. Adriaens T, Baert K, Breyne P, Casaer J, Devisscher S, Onkelinx T, et al. Successful eradication of a suburban Pallas's squirrel Callosciurus erythraeus (Pallas 1779) (Rodentia, Sciuridae) population in Flanders (northern Belgium). *Biol Invasions* (2015) **17**:2517–26.
72. Faulkner KT, Robertson MP, Rouget M, Wilson JR. A simple, rapid methodology for developing invasive species watch lists. *Biol Conserv*. (2014) **179**:25–32. doi: 10.1016/j.biocon.2014.08.014
73. Beauchamp VB, Koontz SM, Suss C, Hawkins C, Kyde KL, Schnase JL. An introduction to Oplismenus undulatifolius (Ard.) Roem. & Schult.(wavyleaf basketgrass), a recent invader in Mid-Atlantic forest understories. *J Torrey Bot Soc*. (2013) **140**:391–413. doi: 10.3159/TORREY-D-13-00033.1
74. Panzacchi M, Cocchi R, Genovesi P, Bertolino S. Population control of coypu Myocastor coypus in Italy compared to eradication in UK: a cost-benefit analysis. *Wildlife Biol*. (2007) **13**:159–71. doi: 10.2981/0909-6396(2007)13[159:PCOCMC]2.0.CO;2
75. Rajmis S, Thiele J, Marggraf RA. Cost-benefit analysis of controlling giant hogweed (Heracleum mantegazzianum) in Germany using a choice experiment approach. *NeoBiota* (2016) **31**:19–41. doi: 10.3897/neobiota.31.8103

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers
in Earth Science

ORIGINAL RESEARCH
published: 25 July 2017
doi: 10.3389/feart.2017.00060

Spatial Downscaling of Alien Species Presences Using Machine Learning

Ioannis N. Daliakopoulos [1,2], Stelios Katsanevakis [3] and Aristides Moustakas [4]*

[1] TM Solutions, Specialized Health and Environmental Services, Crete, Greece, [2] School of Environmental Engineering, Technical University of Crete, Crete, Greece, [3] Department of Marine Sciences, University of the Aegean, Mytilene, Greece, [4] School of Biological and Chemical Sciences, Queen Mary University of London, London, United Kingdom

Edited by:
Yang Liu,
Emory University, United States

Reviewed by:
Yuichi S. Hayakawa,
University of Tokyo, Japan
Tianhai Cheng,
Institute of Remote Sensing and
Digital Earth (CAS), China

***Correspondence:**
Ioannis N. Daliakopoulos
daliakopoulos@tmsolutions.gr

Specialty section:
This article was submitted to
Environmental Informatics,
a section of the journal
Frontiers in Earth Science

Received: 30 April 2017
Accepted: 11 July 2017
Published: 25 July 2017

Citation:
Daliakopoulos IN, Katsanevakis S and
Moustakas A (2017) Spatial
Downscaling of Alien Species
Presences Using Machine Learning.
Front. Earth Sci. 5:60.
doi: 10.3389/feart.2017.00060

Spatially explicit assessments of alien species environmental and socio-economic impacts, and subsequent management interventions for their mitigation, require large scale, high-resolution data on species presence distribution. However, these data are often unavailable. This paper presents a method that relies on Random Forest (RF) models to distribute alien species presence counts at a finer resolution grid, thus achieving spatial downscaling. A bootstrapping scheme is designed to account for sub-setting uncertainty, and subsets are used to train a sufficiently large number of RF models. RF results are processed to estimate variable importance and model performance. The method is tested with an ~8 × 8 km^2 grid containing floral alien species presence and several potentially exploratory indices of climatic, habitat, land use, and soil property covariates for the Mediterranean island of Crete, Greece. Alien species presence is aggregated at 16 × 16 km^2 and used as a predictor of presence at the original resolution, thus simulating spatial downscaling. Uncertainty assessment of the spatial downscaling of alien species' occurrences was also performed and true/false presences and absences were quantified. The approach is promising for downscaling alien species datasets of larger spatial scale but coarse resolution, where the underlying environmental information is available at a finer resolution. Furthermore, the RF architecture allows for tuning toward operationally optimal sensitivity and specificity, thus providing a decision support tool for designing a resource efficient alien species census.

Keywords: downscaling, data analytics, alien species, hydro-ecological data, random forests, vascular plants, Crete

INTRODUCTION

The rate at which species are being translocated by humans beyond their native ranges, through a variety of pathways, has been accelerating (Essl et al., 2015). Alien species pose a grave risk to biodiversity, ecosystem services, and human health, and their presence is an important constituent of the global change that we currently face (Vilà et al., 2011; Simberloff et al., 2013; Katsanevakis et al., 2014), hence there is an urgent need for targeted actions for prevention and mitigation. Despite global efforts to tackle biological invasions, so far there is no sign of saturation in the accumulation of alien species (Hulme et al., 2009; Tittensor et al., 2014; Seebens et al., 2017).

A better understanding of the factors controlling alien species introduction, initial dispersal, establishment success, distribution, abundance, spatio-temporal dynamics, and invasiveness is essential for the efficient prioritization of measures to prevent further introductions and mitigate the impacts of invasive alien species (Byers et al., 2002; Thuiller et al., 2006). Reliable fine scale spatio-temporal information of alien species distribution at large scales is thus crucial (Collingham et al., 2000; Giakoumi et al., 2016). However, the spatial resolution of available data often poses limitations in the analyses. For species distribution data, spatial resolution and spatial extent are typically inversely proportional (Collingham et al., 2000). The European Alien Species Information Network (EASIN; Katsanevakis et al., 2015), which has compiled the largest spatial dataset of alien species distribution in Europe, reports species presence data at a 10 × 10 km^2 spatial resolution and for some species only at country level. Such coarse resolution is often inadequate for the needs of management and research, as data availability up to a point determines the outputs of the analysis in several ways including complexity, generality, utility, and predictive power (Evans et al., 2014; Evans and Moustakas, 2016). Therefore, either more data need to be collected or computational and statistical methods could be used to increase the utility of readily available data and the reliability of the analyses. To that end, accurate methods for downscaling coarse spatial data can be extremely useful in assessments of environmental and socio-economic impacts of alien species and in management interventions for mitigation.

Based on the fundamental assumption that detectable relationships exist between information across spatial scales, spatial downscaling refers to the process and methodologies of using coarse resolution input to infer finer resolution output. Although, it has extensively been used in other scientific disciplines (Trzaska and Schnarr, 2014), downscaling is not a trivial process. Keil et al. (2013) lists four strictly computational approaches that have been used to predict fine-gridded species presence based on a coarse grid. Assumptions made by direct and iterative approaches (Keil et al., 2013) are often criticized for speculating similar species association with environmental variables across scales (Menke et al., 2009). On the other hand, point sampling and clustering approaches make assumptions regarding habitat suitability within the coarse grid (Keil et al., 2013). In their review, Trzaska and Schnarr (2014) distinguish downscaling methods between the relatively straight-forward but normality-limited linear methods (e.g., delta method, simple and multiple linear regression, canonical-correlation analysis, etc.) and the more versatile but data-intensive and extrapolation-limited non-linear methods (e.g., analog method, cluster analysis, artificial neural networks, self-organizing maps, etc.).

Since essentially spatial downscaling is largely based on the inter-relationship between local and large-scale properties, given the absence of fine scale alien species data, potential environmental explanatory covariates available at the resolution of the alien species, as well as at finer resolutions, could be used to infer alien species presences at finer resolutions. While data on alien species presences may be scarce, environmental data may be readily available. Recent advances in remote sensing, social networks, and digital technology resulted in the availability of large spatially and temporally explicit datasets (Moustakas, 2017). Ecology, epidemiology, and biogeography need to employ novel methods for big data analytics combing statistics and computer science, as the analysis of such datasets requires advanced methods for compiling the data, their visualization, and their analyses (Moustakas, 2017; Moustakas and Evans, 2017). Furthermore, computational methods for data analytics and simulation modeling are facilitated by the existence of increased computer power (Moustakas and Evans, 2015).

Recently, methods that generate numerous classifier functions and aggregate their output, widely referred to as "ensembles methods," have attracted wide interest. In this context, Random Forest (RF) algorithms are ensembles of decision trees (Breiman et al., 1984), each trained on a randomly sampled subset of the available dataset, thus reducing the chance of overfitting (Breiman, 2001). In the domain of ecology, RFs have been applied for tropical forest carbon mapping using LiDAR (Light Detection and Ranging)-based carbon estimates (Mascaro et al., 2014), downscaling of global livestock census data (Nicolas et al., 2016), occurrence of fish species in relation to environmental variables (Vezza et al., 2015), forest health and vitality in relation to climate and air pollution parameters (Vitale et al., 2014), classification of tree species using an ensemble of remote sensing data (Naidoo et al., 2012), and vegetation spatial distribution assessment under current and future climate scenarios (Prasad et al., 2006). Especially regarding alien species, RFs have been used for mapping of presence using spatial (Peerbhay et al., 2016) or spatiotemporal (Dorigo et al., 2012) analysis of remote sensing data, prediction of presence based on environmental variables (Cutler et al., 2007; Jarošík et al., 2011), and invasion risk assessment based on biogeographical and life-history variables (Chen et al., 2015).

Here we showcase the applicability of spatial downscaling alien species presences using data from vascular plant species coupled with environmental, potential explanatory, covariates comprised of climatic, soil, habitat, and land use indicators at a finer resolution at the Mediterranean island of Crete, Greece. Apart from investigating the potential of using the fine-resolution environmental covariates as predictors for spatial downscaling alien species presences, the developed methodology also assesses the relevant importance of predictors for the downscaling process as well as visualizing and quantifying their actual response on alien species presences. Furthermore, considering an operational framework for assessing presence, the methodology integrates the use of a detection sensitivity threshold.

CASE STUDY

The Island of Crete

Covering an area of 8,700 km^2, Crete is the largest and most populated island of Greece, and the fifth largest in the Mediterranean. According the Köppen classification, Crete has a Mediterranean—Semiarid climate featuring long and dry summers, and relatively wet and cold winters (Kottek et al., 2006). Crete receives on average about 7.7 billion m^3 of rainfall, of which only and 10–15% produces runoff, while 68–76% evapotranspires and 14–17% infiltrates (Koutroulis et al., 2016). The intense tectonic history has formed the island's complex topography that ranges from sea level to 2,450 m, and is abundant in small, ephemeral watersheds (Tsanis et al., 2011). This highly-rugged terrain has been definitive for human development as well as its spatial allocation (Koutroulis et al., 2016). Similarly, this variability has spurred the development of the wide variety of biotopes present on the island (Sfikas, 1987), ranging from coastal to alpine, and the reciprocal plethora of endemic and rare species that constitute one of the 10 biodiversity hotspots in the Mediterranean (Medail and Quezel, 1997) numbering 1,624 native and 47 introduced vascular floral species (Turland et al., 1993). The importance of this biodiversity is highlighted by the fact that over 30% of the island has been included in the Natura 2000 protected area network (Dimitrakopoulos et al., 2004).

Plant Data

Maps of presence-absence of vascular plant species distributions in Crete were digitized from Turland et al. (1993) and its latest supplement (Chilton and Turland, 2004). The island of Crete and its surrounding islets were divided into 162 grid cells, each covering an area of 8.25×8.25 km^2, following the grid cell size of Turland et al. (1993). On each cell, the native, endemic, and alien species richness was calculated. We used (Turland et al., 1993; Chilton and Turland, 2004; and references therein) to define native (nnat = 1,395) and endemic (nend = 174) species, and the vascular plants from D'Agata et al. (2009) that are listed in Chilton and Turland (2004) and Turland et al. (1993) were used to define alien species richness. Only species present in at least two cells were used (nalien = 47). Coarse-grid information was estimated by aggregating this dataset by a factor of two, thus reducing resolution to grid cells of 16.5×16.5 km^2. The spatial distribution of the original as well as the resampled data regarding alien species presences are visualized in **Figure 1**. All input variables and their ranges (min – max values within each cell) are listed in **Table 1**.

Habitat Data

Habitat classification relied on the most detailed resolution available of the CORINE Landcover (level 3, spatial resolution 100 m; EEA-ETC/TE., 2002), to calculate the richness and percentage of every land cover class within every grid cell, using Patch Analyst 5.1 within ArcGIS. In order to avoid potential temporal deviance between habitat classifications and species presences in cells, the last updated available supplement for the flora of Crete published in 2008 (Chilton and Turland, 2004) and the closest available time snapshot of the CORINE landcover for Crete in 2010 were used. The classification process resulted in 29 habitat types, of which 9 agricultural, 7 artificial, and 13 natural. We recorded habitat richness per cell as the number of different land cover types present on each cell (total, artificial, agricultural, and natural habitat richness) as well as percentage of cover (total, artificial, agricultural, and natural % of cell cover).

Climatic, Soil, and Altitude Data

Climatic variables were derived from WorldClim (Hijmans et al., 2005) for Crete and surrounding islets. The original resolution of the climatic data was 1×1 km^2. In order to re-scale them to 8.25 km and match them with the grid of the plant data, the mean values of the 1 km data within the 8.25 km cells were calculated and used. The climatic variables used were annual mean temperature (Tempmean), annual mean temperature of warmest quarter (Tempwarm), annual mean temperature of coldest quarter (Tempcold), all in °C, annual mean precipitation (Precipmean), precipitation of wettest quarter (Precipwet), and precipitation of driest quarter (Precipdry), all in mm year^{-1}. Soil data were derived from SoilGrid (Hengl et al., 2014) and rescaled from 1 to 8.25 km as the climatic data. The soil variable used was soil richness in the cell (Soildiv) derived as the number of different soil types occurring within each cell. The indices of elevation recorded were the mean of all elevation values within the cell (Alt) and the range of elevation within the cell (Alt range) both in meters.

METHODOLOGY

Random Forests

Random Forests (RFs; Breiman, 2001) take advantage of boosting (Schapire et al., 1998) and bagging (bootstrap aggregating; Breiman, 1996a) of the Classification And Regression Tree (CART; Breiman et al., 1984) model, and adapt a more random but nevertheless more efficient node splitting strategy than standard CARTs (Liaw and Wiener, 2002). In RFs, each individual tree is developed after the following steps: (1) Given a set of training data N, n random samples with repetition (bootstrap) are taken as training set; (2) For each node of the tree, M input variables are determined, and $m \ll M$, variables are selected for each node. The most important variable randomly chosen is used as a node. The value of m remains constant; (3)Each tree is developed to its maximum expansion.

RFs have been employed in a wide variety of classification and prediction problems (Scornet et al., 2015; Cano et al., 2017) as they are among the most effective computationally-intensive algorithms to extract information from unstable estimates (Scornet et al., 2015). They are especially well suited for large, high-dimensional datasets, where problem complexity and scale render direct discovery of a good model in a single step impossible (Büchlmann and Yu, 2002; Kleiner et al., 2014; Wager et al., 2014). The fact that RFs require tuning of only two parameters (the tree population in each forest and the number of input variables m randomly selected at each node) for which they are usually not very sensitive (Liaw and Wiener, 2002), and their accuracy and competence when faced with scarce, multivariate datasets of intricate structure (Scornet et al., 2015), have greatly contributed to their popularity.

Similar to other data-driven approaches, RFs may not perform equally well when the task at hand is extrapolatory beyond the range of the recovered predictor-predictand relationship or involves scenario analysis (Daliakopoulos and Tsanis, 2016). Furthermore, Strobl and Boulesteix (2007) showed that variable importance measures of the original RF algorithm may be biased due to differences among predictor structure and scale, adding to the interpretability challenges of data-driven methods. Nevertheless, an extensive data-driven model inter-comparison by Fernández-Delgado et al. (2014) showed that they may be the first weapon of choice for real-world problems.

Evaluation Criteria

Typically, CARTs error is estimated following the out-of-bag (OOB) error $R(D)$ of a selection of the input observations based on bagging, otherwise an OOB sample D (James et al., 2013).

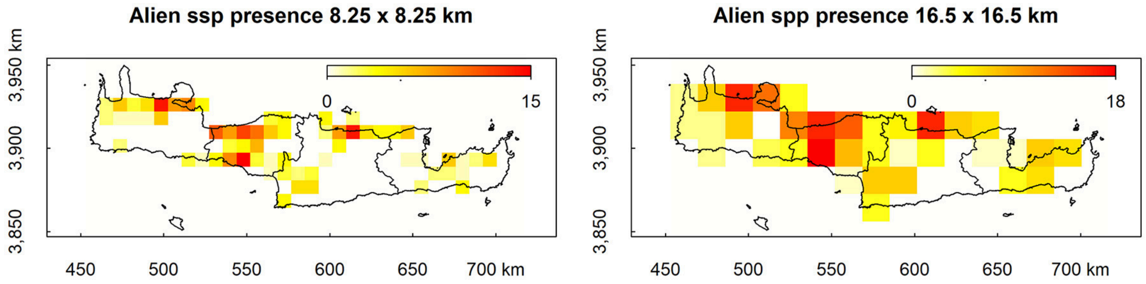


FIGURE 1 | Original **(Left)** and resampled **(Right)** spatial distribution of alien species presence in Crete.

TABLE 1 | Environmental variables used as input for the estimation of alien species presence.

Type	Variable	Value range
Ecological	Habitat richness	2–16
	Artificial habitat richness	0–9
	Agricultural habitat richness	0–7
	Natural habitat richness	1–9
	Endemic species richness	0–54
Topographical	Altitude	0–2,673 m
	Altitude range	0–1,579 m
Climate	Mean annual temperature	0–19.0°C
	Maximum annual temperature	17.1–26.0°C
	Minimum annual temperature	3.2–13.2°C
	Temperature range	11.9–13.9°C
	Mean annual precipitation	519–1,023 mm
	Maximum annual precipitation	265–539 mm
	Minimum annual precipitation	0–31 mm
	Precipitation range	260–514 mm
Land use	Natura % of cover	0–100%
	Artificial % of cover	0–43%
	Agricultural % of cover	0–100%
	Natural % of cover	0–100%
	Soil type richness	1–8

In RFs, for each tree t, prediction error of D is estimated before and after randomly permuting the values of the j-th variable, thus giving $R\left(D_n^t\right)$ and $R\left(D_n^{tj}\right)$, respectively. Typically, imbalanced datasets favor correct classification of the majority class, nevertheless, RFs can account for this bias by adjusting the voting cut-off from the default $1/c$, where c is the number of classes. This provides additional flexibility to the RF algorithm (Ma et al., 2006) and allows for favoring sensitivity or specificity to different classes. A variable can be considered a strong predictor when permuting it increases the prediction error (Gregorutti et al., 2017), therefore it's importance I_V can be defined as:

$$I_V\left(X_j,\, t\right) = R\left(D_n^{tj}\right) - R\left(D_n^t\right) \tag{1}$$

The Mean Decrease in Accuracy (MDA) is estimated by averaging this difference over all trees, and normalizing it by the standard deviation of the differences. The more the accuracy of the RF decreases due to the exclusion (or permutation) of a single predictor, the more important that predictor is considered, and therefore variables with a large MDA are more important for data classification.

Gini is one of the most encountered impurity functions, providing a measure of the "goodness-of-split" for CARTs by favoring splits that allocate a single pure node for the largest class and the rest for the remaining classes (Breiman, 1996b). The Gini index for a node t can be calculated as:

$$I_G\left(t\right) = \sum_{i \neq j}^{c} p\left(i|t\right) p\left(j|t\right) = 1 - \sum_{j}^{c} p\left(j|t\right)^2 \tag{2}$$

where c is the number of classes and $p\left(i|t\right)$, $p\left(j|t\right)$ are the estimated probabilities of classes i, j at node t (Cano et al., 2017). In this context, Mean Decrease Gini (MDG) aggregates the Gini gain over all splits and trees to assess the classifying capacity of a variable (Friedman et al., 2009) and is thus a metric of the homogeneity of nodes and leaves in the RF (Bluemke and Stepień, 2016).

MDA and MDG can rank each independent variable for its effectiveness as a predictor of alien species richness, but don't show or quantify the actual positive, negative, humped, etc. relationship between them. Nevertheless, this is an elementary process under conditions of multiple acting variables (Häring et al., 2012), such as cumulative human impacts. For this reason, partial dependence plots (Friedman, 2001; Friedman et al., 2009) can be used to depict the relationship of alien species presence probability on each predictor after averaging out the effects of all classification predictors (Cutler et al., 2007).

Finally, the Receiver Operating Characteristics (ROC) analysis has been an indispensable tool for signal detection and diagnostic systems. As documented by Pontius and Si (2014), ROC has been employed in a wide range of applications requiring a threshold-independent measure to compare predicted against observed values. ROC plots have been previously considered in plant ecology, both at a theoretical (Guisan and Zimmermann, 2000) and applied (Manel et al., 2002; Wang et al., 2014) level as effective indicators of model performance independent of the threshold probability. Typically, ROC curves depict true positive rate (*TPR*), otherwise sensitivity, recall or hit rate, against true negative rate (*TNR*), otherwise called specificity. In terms of model estimates, *TPR* and *TNR* are defined as:

$$TPR = {}^{TP}\!/_{P} = \frac{TP}{TP + FN} \quad (3)$$

$$TNR = {}^{TN}\!/_{N} = \frac{TN}{TN + FP} \quad (4)$$

where T, F, P, and N stand for true, false, positive and negative, respectively. The complementary values of *TPR* and *TNR* are false negative rate (*FNR*), otherwise miss rate, and false positive rate (*FPR*), otherwise fall-out or false alarm. Based on these values, the Matthew's correlation coefficient (MCC; Matthews, 1975), a reduction of the Pearson correlation coefficient for binary variables (Baldi and Brunak, 2001), is a popular evaluation criterion of machine learning performance (Bhasin and Raghava, 2004; Chen et al., 2004; Bao and Cui, 2005):

$$MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TN + FN)\,(TN + FP)\,(TP + FN)\,(TP + FP)}} \quad (5)$$

MCC has an advantage in imbalanced datasets where the disparity in the number of presence and absence samples is significant.

Random Forest Application

Experiments were developed using the latest (v4.6–12) implementation of Breiman and Cutler's original Fortran code by Liaw and Wiener (2002) in R. While RFs can be trained very efficiently and avoid overfitting (Breiman, 2001), predictions and variable significance ranking are seldom the identical after each random training, especially for small datasets. To account for this uncertainty, a bootstrapping approach of training multiple RFs is adopted. For each training iteration k, RF_k is presented with 70% of the dataset, sampled with replacement, and the remaining is reserved for testing. Presenting only part of the dataset to the RFs also simulates operational use where only part of the study area is sampled at fine grid and the rest is sampled at coarse-grid resolution. Furthermore, as subsets of alien species presence and absence were imbalanced, training was executed using a variable training cutoff, ranging from 0.1 to 0.9. The full code in R used for the analysis is provided in Supplementary Material.

RESULTS

Importance and Gini

Mean decrease in accuracy (MDA) results as estimated from bootstrap randomizations indicate that, apart from the coarse resolution alien species presence, the percentage of natural cover within each cell was the most important predictor of alien species presence, followed by the endemic species richness, altitude, minimum temperature, and altitude range within each cell (**Figure 2**, left). From the ones explored here, the least predictive in MDA were artificial habitat richness, temperature range, habitat richness, the percentage of the surface area of each cell within the Natura 2,000 protected area network, and the soil type richness (**Figure 2**, left). In the latter cases, some bootstrap samples have yielded negative results suggesting that permuting these variables from the predictor vector increases accuracy. Results in Mean Decrease Gini (MDG) are in general agreement with those of MDA, also evaluating natural cover, endemic species richness, and altitude as the most efficient splitting variables (**Figure 2**, right). Agricultural cover replaces minimum temperature for the MDG rating but both variables score highly for both criteria. The least efficient node splits according to MDG were performed by artificial habitat richness, natural habitat richness, agricultural habitat richness, soil type richness, and temperature range (**Figure 2**, right). Emphatically, artificial habitat richness is the worst predictor for both metrics, essentially boosting the noise in the dataset.

Partial Dependence Plots

Results from partial dependence plots among the most predictive variables according to the MDA and MDG criteria indicate that the percentage of natural cover has an overall positive relationship with alien species richness, while the percentage of agricultural cover has an overall negative relationship with alien species richness (**Figure 3**). Altitude, and altitude range has an overall positive relationship between alien species richness, mean temperature has a negative relationship for larger temperature values while mean annual precipitation and precipitation range has a humped relationship with alien species richness (**Figure 3**). Therefore, for the case at hand, in the event of a survey priority may be given to low-temperature, elevated natural areas with high topographic variability, far from agricultural use and precipitation extremes. The percentage of each cell within the Natura 2,000 protected area network has a positive relationship with alien species richness (**Figure 3**), albeit this variable was not within the most predictive of alien species richness based on MDA or MDG.

As shown by the results, the bootstrapping method followed herein is helpful for drawing a more robust conclusion, particularly regarding the partial dependence plots. Bootstrapped predictors (solid black lines in **Figure 3**) are more stable, less prone to overfit and more inclusive than single experiment predictors. This becomes obvious in the Mean Temperature plot of **Figure 3**, where the red line representing an OOB sample does not cover the entire range of temperature values in the dataset. As low temperatures are not common in the dataset, the OOB estimation of dependence does not always include these values. Using an additional layer of bootstrapping ensures that the full range of values is explored.

Uncertainly and Risk Assessment

True negative detection rates (TNR; not detecting alien species in cells where alien species are not present) declines with an increasing cut-off rate while true positive detection rates (TPR; detecting alien species in cells where alien species are present) increases with an increasing cut-off rate (**Figure 4**). MCC values indicate a strong positive relationship at cut-offs between 0.2 and 0.5 and are otherwise acceptable correlation. When cut-off increases TNs decrease and TFs increase, therefore more alien species can be detected but by being more exhaustive more false

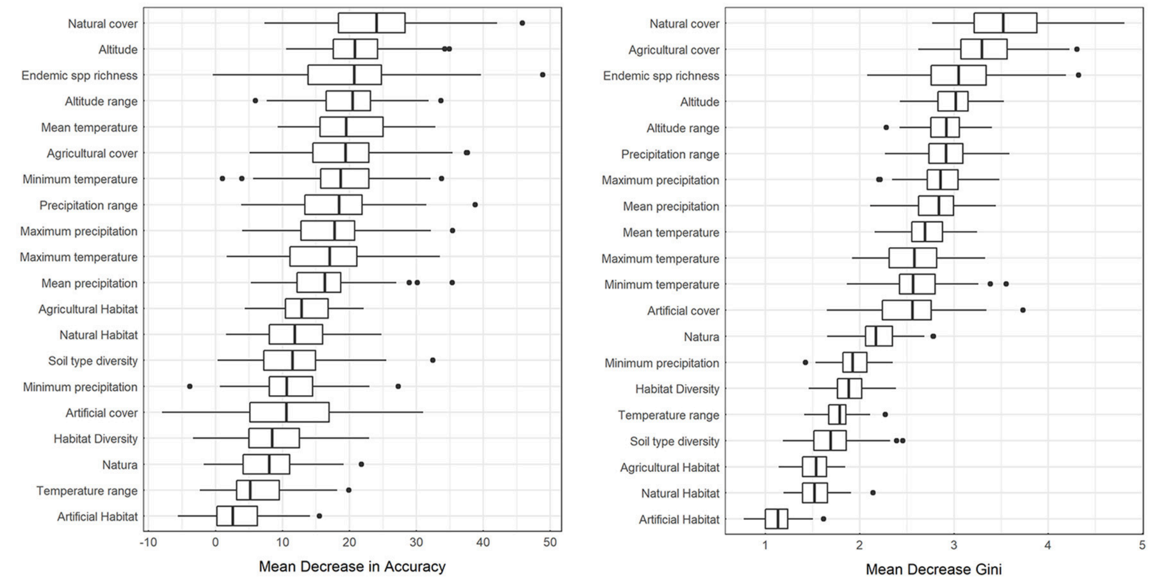


FIGURE 2 | Distribution of Mean Decrease in Accuracy (MDA) and Mean Decrease Gini (MDG) estimated from the bootstrap runs.

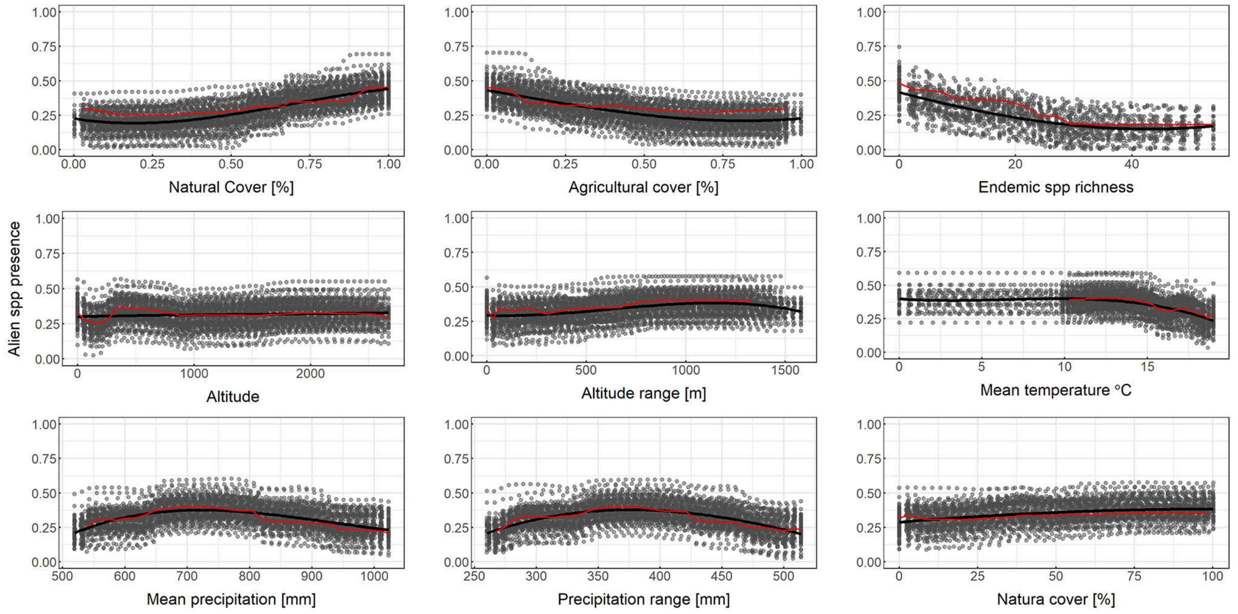


FIGURE 3 | Partial dependence plots for selected independent variables for random forest downscaling of alien species presence. Y-axis is on the logit scale. Here 3rd degree polynomial splines (solid black lines) are fitted over the output the Monte Carlo runs (gray points). Red lines connect points from a single random sample of the bootstrap experiment.

alarms are also generated. When cut-off remains low, less risk is taken with surveying resources but a significant fraction of alien species presences is missed.

DISCUSSION

Aichi Target 9 of the Convention on Biological Diversity, states that "by 2020, invasive alien species and pathways are identified and prioritized, priority species are controlled or eradicated and measures are in place to manage pathways to prevent their introduction and establishment." Prioritization of species, pathways of introduction, and sites for management measures is crucial for the implementation of Aichi Target 9, but the lack of adequate data often compromises the ability of countries to make substantial progress (McGeoch et al., 2016). Large scale, high-resolution data on alien species distributions as well as the

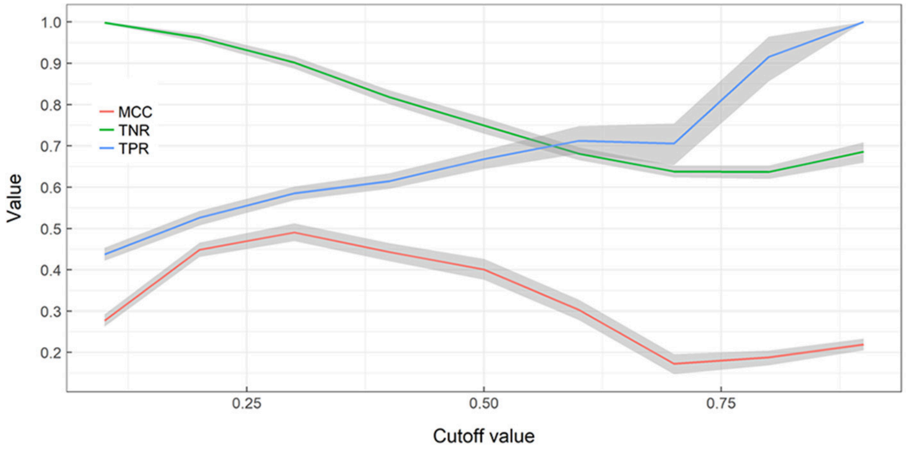


FIGURE 4 | Performance of Random Forest ensemble vs. training cut-off value. A nonparametric bootstrap is used to obtaining confidence limits (gray areas) and bootstrap means for the Matthews correlation coefficient (MCC), True Negative Rate (TNR), True Positive Rate (TPR), without assuming normality.

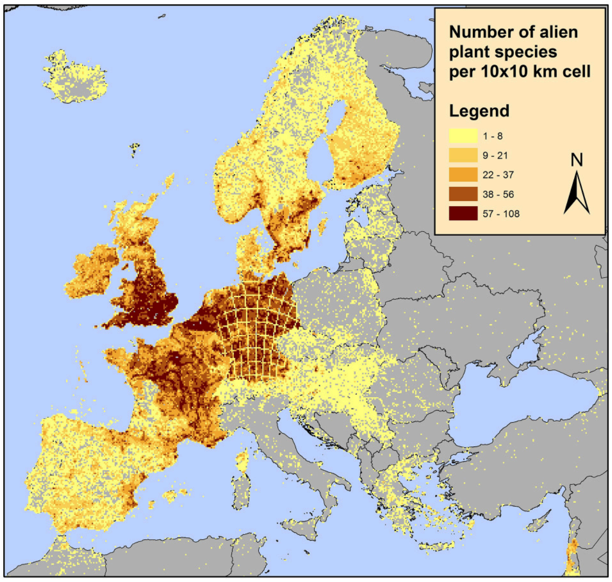


FIGURE 5 | Distribution of alien plants in Europe on a 10 × 10 km^2 grid according to the available data in the European Alien Species Information Network (EASIN; Katsanevakis et al., 2015). These spatial data, integrated in EASIN, originate from the following sources: (1) the Global Biodiversity Information Facility (GBIF; http://www.gbif.org/); (2) the Global Invasive Species Information Network (GISIN; http://www.gisin.org); (3) the Regional Euro-Asian Biological Invasions Centre (REABIC; http://www.reabic.net/); (4) the European and Mediterranean Plant Protection Organization (EPPO; http://www.eppo.int/); (5) the Norwegian Biodiversity Information Centre (NBIC, http://www.biodiversity.no/) and (6) EASIN-Lit (http://easin.jrc.ec.europa.eu/About/EASIN-Lit; Trombetti et al., 2013).

associated human and environmental pressures are necessary when performing a spatially explicit quantitative environmental and socio-economic evaluation and prioritizing interventions for their mitigation and management (Hobbs and Humphries, 1995; McGeoch et al., 2016).

It is only evident that substantial part of model output reliability is based on model input validity, thus uncertainty needs to be accounted for (Burgman et al., 2005). Therefore, investment in conservation actions that have been supported by poor field observations has a high probability of yielding poor outcomes (McGeoch et al., 2016), regardless of the subsequent decision process quality. Moreover, ecological processes are often inherently non-linear, and potential explanatory covariates include correlated independent variables, as well as interacting effects. As shown here, RFs can make use of input variables without prior scaling and knowledge of physical or other dependences between predictors and predictands. RFs make no assumptions regarding linearity, handle multiple correlated independent variables well, quantify the importance of each predictor variable, and through partial plots depict the contribution of each independent variable. By assessing the importance of predictors for the desired classification, RFs can effectively permute noisy or otherwise unprofitable data. In addition to enhancing existing model accuracy, this output can have operational value by providing data/survey managers with hints about which data recovery is worth investing in and which not.

Decision makers' requirements for confronting environmental risks and prioritizing mitigation measures at fine grid scale are often much higher than what model limitations and data availability allow. In these cases, a commonly used approach is to employ statistical tools in order to infer impacts at the required scale (Trzaska and Schnarr, 2014). It is crucial to identify and evaluate the premises under which analyses and techniques are used to deduce such output, and to recognize their constraints and inherent uncertainties. In the case of alien species presence downscaling, the approach relies on the assumption that fine-resolution presence is a combination of a coarse-grid presence assessment and environmental conditions, and fine-grid environmental conditions. A common drawback of such approaches is that inherent uncertainties from both initial projections and downscaling procedure are not quantified or adequately conveyed to decision makers and end-users, thus creating an over-confidence to the inferred results and causing validation and updating of downscaled information to be omitted.

Here we have performed spatial downscaling of alien species presences using a relatively idiosyncratic and tricky dataset: the spatial distribution of alien species is clustered, the spatial sample size in terms of the number of cells of the grid of the study area are limited (162 cells in total), and the study area is an island meaning that there are edge effects, unequal land surface areas in coastal cells than in mainland cells, and a very idiosyncratic physical geography, as the island has over 50 mountain summits above 2,000 m (Vogiatzakis et al., 2003). Despite this, the method worked well in the sense that environmental data/covariates of finer scale than the ones of alien species presences can produce finer resolution alien species presences spatial data, and predicted presences or absences were verified and thus the predictive accuracy is explicitly quantified. While additional validation studies in different spatial contexts may highlight other downscaling determining variables, this study outlines an exploratory analysis for variable selection and operational use where underlying environmental information is available at higher resolution. In view of new, spatially and temporally richer data sources (e.g., remote sensing products), results of the present study can be greatly enhanced. Starting from a cost-effective targeted survey design based on the proposed downscaling approach, an improved alien species mapping result can be reached. Beyond the downscaling process itself, a better understanding of alien species distribution and environmental factors that facilitate their presence on the island can be achieved.

Furthermore, the RF architecture allows for tuning toward operationally optimal sensitivity and specificity, thus providing a decision support tool for designing a resource-efficient alien species census. For example, according to one of the most updated alien species dataset in Europe, the distribution of alien plant species appears to be highly clustered with some countries such as the UK, Germany and France appearing to contain the majority of alien species (EASIN dataset; see **Figure 5** and references therein). This is unlikely to reflect the actual situation; alien species sampling effort is not evenly distributed among countries and even within countries some areas are better sampled than others. Using the approach proposed here, areas where alien species are not detected but are likely to occur and thus detected once sampled as well as areas where alien species are not detected but are unlikely to occur once sampled can be identified. Additionally, the acceptable risk of false negative and false positive occurrences, also reflecting field detection effort and human labor, can be quantified. In this study, the predicted variable was alien species richness of all alien species, however, given the number of alien species records in the EASIN dataset, the analysis performed here can be adapted at single species level.

CONCLUSIONS

The science needs for conducting research on biological invasions and the policy needs for management prioritization to prevent further introductions and to mitigate the impacts of invasive alien species, include high-resolution spatiotemporal data of species distributions. We herein demonstrated the applicability of RFs for spatial downscaling, which is an effective, advantageous and useful approach when environmental data are available at better resolution than that of alien species' spatial information. In relation to other downscaling approaches, RFs don't rely on assumptions about environmental parameters and their effect on alien species presence; rather these relationships emerge from the classification process. This way, RFs can provide a better understanding of facilitating and limiting factors of alien species presence, both for research and management purposes. By effectively downscaling coarse-grid alien presence, the RFs can facilitate targeted actions for prevention and mitigation, thus providing an operational exploration tool.

AUTHOR CONTRIBUTIONS

ID developed the methodology and code, and analyzed the results, SK framed the work within the international context and analyzed results, and AM had the idea, contributed the data, and analyzed the results. All authors contributed equally to the writing process.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: http://journal.frontiersin.org/article/10.3389/feart.2017.00060/full#supplementary-material

REFERENCES

Baldi, P., and Brunak, S. (2001). *Bioinformatics : The Machine Learning Approach*. MIT Press.

Bao, L., and Cui, Y. (2005). Prediction of the phenotypic effects of non-synonymous single nucleotide polymorphisms using structural and evolutionary information. *Bioinformatics* 21, 2185–2190. doi: 10.1093/bioinformatics/bti365

Bhasin, M., and Raghava, G. P. S. (2004). ESLpred: SVM-based method for subcellular localization of eukaryotic proteins using dipeptide composition and PSI-BLAST. *Nucleic Acids Res.* 32, W414–W419. doi: 10.1093/nar/gkh350

Bluemke, I., and Stepień, A. (2016). *Selection of Metrics for the Defect Prediction*. Cham: Springer.

Breiman, L. (1996a). Bagging predictors. *Mach. Learn.* 24, 123–140. doi: 10.1007/BF00058655

Breiman, L. (1996b). Technical note: some properties of splitting criteria. *Mach. Learn.* 24, 41–47. doi: 10.1007/BF00117831

Breiman, L. (2001). Random forests. *Mach. Learn.* 45, 5–32. doi: 10.1023/A:1010933404324

Breiman, L., Friedman, J., Stone, C., and Olshen, R. (1984). *Classification and Regression Trees*. New York, NY: CRC Press.

Büchlmann, P., and Yu, B. (2002). Analyzing bagging. *Ann. Stat.* 30, 927–961. doi: 10.1214/aos/1031689014

Burgman, M. A., Lindenmayer, D. B., and Elith, J. (2005). Managing landscapes for conservation under uncertainty. *Ecology* 86, 2007–2017. doi: 10.1890/04-0906

Byers, J. E., Reichard, S., Randall, J. M., Parker, I. M., Smith, C. S., Lonsdale, W. M., et al. (2002). Directing research to reduce the impacts of nonindigenous species. *Conserv. Biol.* 16, 630–640. doi: 10.1046/j.1523-1739.2002.01057.x

Cano, G., Garcia-Rodriguez, J., Garcia-Garcia, A., Perez-Sanchez, H., Benediktsson, J. A., Thapa, A., et al. (2017). Automatic selection of molecular descriptors using random forest: application to drug discovery. *Expert Syst. Appl.* 72, 151–159. doi: 10.1016/j.eswa.2016.12.008

Chen, L., Peng, S., and Yang, B. (2015). Predicting alien herb invasion with machine learning models: biogeographical and life-history traits both matter. *Biol. Invasions* 17, 2187–2198. doi: 10.1007/s10530-015-0870-y

Chen, Y. C., Lin, Y. S., Lin, C. J., and Hwang, J. K. (2004). Prediction of the bonding states of cysteines using the support vector machines based on multiple feature vectors and cysteine state sequences. *Proteins Struct. Funct. Bioinform.* 55, 1036–1042. doi: 10.1002/prot.20079

Chilton, L., and Turland, N. (2004). *Flora of Crete: Supplement, I. I., Additions 1997-2004*. Available onlie at: Publ. internet http://www.marengowalks

Collingham, Y. C., Wadsworth, R. A., Huntley, B., and Hulme, P. E. (2000). Predicting the spatial distribution of non-indigenous riparian weeds: issues of spatial scale and extent. *J. Appl. Ecol.* 37, 13–27. doi: 10.1046/j.1365-2664.2000.00556.x

Cutler, D. R., Edwards, T. C., Beard, K. H., Cutler, A., Hess, K. T., Gibson, J., et al. (2007). Random forests for classification in ecology. *Ecology* 88, 2783–2792. doi: 10.1890/07-0539.1

D'Agata, C., Skoula, M., and Brundu, G. (2009). A preliminary inventory of the alien flora of Crete (Greece). *Bocconea* 23, 301–315.

Daliakopoulos, I. N., and Tsanis, I. K. (2016). Comparison of an artificial neural network and a conceptual rainfall–runoff model in the simulation of ephemeral streamflow. *Hydrol. Sci. J.* 61, 2763–2774. doi: 10.1080/02626667.2016.1154151

Dimitrakopoulos, P. G., Memtsas, D., and Troumbis, A. Y. (2004). Questioning the effectiveness of the natura 2000 special areas of conservation strategy: the case of crete. *Glob. Ecol. Biogeogr.* 13, 199–207. doi: 10.1111/j.1466-822X.2004.00086.x

Dorigo, W., Lucieer, A., Podobnikar, T., and Čarni, A. (2012). Mapping invasive *Fallopia japonica* by combined spectral, spatial, and temporal analysis of digital orthophotos. *Int. J. Appl. Earth Obs. Geoinf.* 19, 185–195. doi: 10.1016/j.jag.2012.05.004

EEA-ETC/TE. (2002). *CORINE Land Cover Update. I& CLC2000 Project*.

Essl, F., Bacher, S., Blackburn, T. M., Booy, O., Brundu, G., Brunel, S., et al. (2015). Crossing frontiers in tackling pathways of biological invasions. *Bioscience* 65, 769–782. doi: 10.1093/biosci/biv082

Evans, M. R., Benton, T. G., Grimm, V., Lessells, C. M., O'Malley, M. A., Moustakas, A., et al. (2014). Data availability and model complexity, generality, and utility: a reply to Lonergan. *Trends Ecol. Evol.* 29, 302–303. doi: 10.1016/j.tree.2014.03.004

Evans, M. R., and Moustakas, A. (2016). A comparison between data requirements and availability for calibrating predictive ecological models for lowland UK woodlands: learning new tricks from old trees. *Ecol. Evol.* 6, 4812–4822. doi: 10.1002/ece3.2217

Fernández-Delgado, M., Cernadas, E., Barro, S., and Amorim, D. (2014). Do we need hundreds of classifiers to solve real world classification problems? *J. Mach. Learn. Res.* 15, 3133–3181. Available online at: http://jmlr.org/papers/v15/delgado14a.html

Friedman, J. (2001). Greedy function approximation: a gradient boosting machine. *Ann. Stat.* 29, 1189–1232. doi: 10.1214/aos/1013203451

Friedman, J., Hastie, T., and Tibshirani, R. (2009). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. New York, NY: Springer-Verlag.

Giakoumi, S., Guilhaumon, F., Kark, S., Terlizzi, A., Claudet, J., Felline, S., et al. (2016). Space invaders; biological invasions in marine conservation planning. *Divers. Distrib.* 22, 1220–1231. doi: 10.1111/ddi.12491

Gregorutti, B., Michel, B., and Saint-Pierre, P. (2017). Correlation and variable importance in random forests. *Stat. Comput.* 27, 659–678. doi: 10.1007/s11222-016-9646-1

Guisan, A., and Zimmermann, N. E. (2000). Predictive habitat distribution models in ecology. *Ecol. Modell.* 135, 147–186. doi: 10.1016/S0304-3800(00)00354-9

Häring, T., Dietz, E., Osenstetter, S., Koschitzki, T., and Schröder, B. (2012). Spatial disaggregation of complex soil map units: a decision-tree based approach in Bavarian forest soils. *Geoderma* 185–186, 37–47. doi: 10.1016/j.geoderma.2012.04.001

Hengl, T., de Jesus, J. M., MacMillan, R. A., Batjes, N. H., Heuvelink, G. B. M., Ribeiro, E., et al. (2014). SoilGrids1km — global soil information based on automated mapping. *PLoS ONE* 9:e105992. doi: 10.1371/journal.pone.0105992

Hijmans, R. J., Cameron, S. E., Parra, J. L., Jones, P. G., and Jarvis, A. (2005). Very high resolution interpolated climate surfaces for global land areas. *Int. J. Climatol.* 25, 1965–1978. doi: 10.1002/joc.1276

Hobbs, R. J., and Humphries, S. E. (1995). An integrated approach to the ecology and management of plant invasions. *Conserv. Biol.* 9, 761–770. doi: 10.1046/j.1523-1739.1995.09040761.x

Hulme, P. E., Pyšek, P., Nentwig, W., and Vilà, M. (2009). Will threat of biological invasions unite the european union? *Science* 324, 40–41. doi: 10.1126/science.1171111

James, G., Witten, D., Hastie, T., and Tibshirani, R. (2013). *An Introduction to Statistical Learning, Springer Texts in Statistics*. New York, NY: Springer.

Jarošík, V., Pyšek, P., Foxcroft, L. C., Richardson, D. M., Rouget, M., and MacFadyen, S. (2011). Predicting incursion of plant invaders into kruger national park, South Africa: the interplay of general drivers and species-specific factors. *PLoS ONE* 6:e28711. doi: 10.1371/journal.pone.0028711

Katsanevakis, S., Deriu, I., D'Amico, F., AL, N., and S, P.S. (2015). European alien species information network (EASIN): supporting European policies and scientific research. *Manage. Biol. Invasions* 6, 147–157. doi: 10.3391/mbi.2015.6.2.05

Katsanevakis, S., Wallentinus, I., Zenetos, A., Leppäkoski, E., Çinar, M. E., Oztürk, B., et al. (2014). Impacts of invasive alien marine species on ecosystem services and biodiversity: a pan-European review. *Aquat. Invasions* 9, 391–423. doi: 10.3391/ai.2014.9.4.01

Keil, P., Belmaker, J., Wilson, A. M., Unitt, P., and Jetz, W. (2013). Downscaling of species distribution models: a hierarchical approach. *Methods Ecol. Evol.* 4, 82–94. doi: 10.1111/j.2041-210x.2012.00264.x

Kleiner, A., Talwalkar, A., Sarkar, P., and Jordan, M. I. (2014). A scalable bootstrap for massive data. *J. R. Stat. Soc. Ser. B Stat. Methodol.* 76, 795–816. doi: 10.1111/rssb.12050

Kottek, M., Grieser, J., Beck, C., Rudolf, B., and Rubel, F. (2006). World map of the Köppen-Geiger climate classification updated. *Meteorol. Zeitschrift* 15, 259–263. doi: 10.1127/0941-2948/2006/0130

Koutroulis, A., Grillakis, M., Daliakopoulos, I., Tsanis, I., and Jacob, D. (2016). Cross sectoral impacts on water availability at +2°C and +3°C for east mediterranean island states: the case of crete. *J. Hydrol.* 532, 16–28. doi: 10.1016/j.jhydrol.2015.11.015

Liaw, A., and Wiener, M. (2002). Classification and regression by randomForest. *R News* 2, 18–22. Available online at: http://www.bios.unc.edu/~dzeng/BIOS740/randomforest.pdf

Ma, Y., Guo, L., and Cukic, B. (2006). "A statistical framework for the prediction of fault-proneness," in *Advances in Machine Learning Application,* eds D. Zhang and J. J.-P. Tsai (Hershey, PA: Idea Group Pub), 480.

Manel, S., Williams, H. C., and Ormerod, S. J. (2002). Evaluating presence-absence models in ecology: the need to account for prevalence. *J. Appl. Ecol.* 38, 921–931. doi: 10.1046/j.1365-2664.2001.00647.x

Mascaro, J., Asner, G. P., Knapp, D. E., Kennedy-Bowdoin, T., Martin, R. E., Anderson, C., et al. (2014). A Tale of Two "Forests": random forest machine learning aids tropical forest carbon mapping. *PLoS ONE* 9:e85993. doi: 10.1371/journal.pone.0085993

Matthews, B. W. (1975). Comparison of the predicted and observed secondary structure of T4 phage lysozyme. *Biochim. Biophys. Acta Protein Struct.* 405, 442–451. doi: 10.1016/0005-2795(75)90109-9

McGeoch, M. A., Genovesi, P., Bellingham, P. J., Costello, M. J., McGrannachan, C., and Sheppard, A. (2016). Prioritizing species, pathways, and sites to achieve conservation targets for biological invasion. *Biol. Invasions* 18, 299–314. doi: 10.1007/s10530-015-1013-1

Medail, F., and Quezel, P. (1997). Hot-spots analysis for conservation of plant biodiversity in the mediterranean basin. *Ann. Missouri Bot. Gard.* 84, 112–127. doi: 10.2307/2399957

Menke, S. B., Holway, D. A., Fisher, R. N., and Jetz, W. (2009). Characterizing and predicting species distributions across environments and scales: argentine ant occurrences in the eye of the beholder. *Glob. Ecol. Biogeogr.* 18, 50–63. doi: 10.1111/j.1466-8238.2008.00420.x

Moustakas, A. (2017). Spatio-temporal data mining in ecological and veterinary epidemiology. *Stoch. Environ. Res. Risk Assess.* 31, 829–834. doi: 10.1007/s00477-016-1374-8

Moustakas, A., and Evans, M. R. (2015). Coupling models of cattle and farms with models of badgers for predicting the dynamics of bovine tuberculosis

(TB). *Stoch. Environ. Res. Risk Assess.* 29, 623–635. doi: 10.1007/s00477-014-1016-y

Moustakas, A., and Evans, M. R. (2017). A big-data spatial, temporal and network analysis of bovine tuberculosis between wildlife (badgers) and cattle. *Stoch. Environ. Res. Risk Assess.* 31, 315–328. doi: 10.1007/s00477-016-1311-x

Naidoo, L., Cho, M. A., Mathieu, R., and Asner, G. (2012). Classification of savanna tree species, in the Greater Kruger National Park region, by integrating hyperspectral and LiDAR data in a Random Forest data mining environment. *ISPRS J. Photogramm. Remote Sens.* 69, 167–179. doi: 10.1016/j.isprsjprs.2012.03.005

Nicolas, G., Robinson, T. P., Wint, G. R. W., Conchedda, G., Cinardi, G., and Gilbert, M. (2016). Using random forest to improve the downscaling of global livestock census data. *PLoS ONE* 11:e0150424. doi: 10.1371/journal.pone.0150424

Peerbhay, K., Mutanga, O., Lottering, R., and Ismail, R. (2016). Mapping Solanum mauritianum plant invasions using WorldView-2 imagery and unsupervised random forests. *Remote Sens. Environ.* 182, 39–48. doi: 10.1016/j.rse.2016.04.025

Pontius, R. G. Jr, and Si, K. (2014). The total operating characteristic to measure diagnostic ability for multiple thresholds. *Int. J. Geogr. Inf. Sci.* 28, 570–583. doi: 10.1080/13658816.2013.862623

Prasad, A. M., Iverson, L. R., and Liaw, A. (2006). Newer classification and regression tree techniques: bagging and random forests for ecological prediction. *Ecosystems* 9, 181–199. doi: 10.1007/s10021-005-0054-1

Schapire, R. E., Freund, Y., Bartlett, P., and Lee, W. S. (1998). Boosting the margin: a new explanation for the effectiveness of voting methods. *Ann. Stat.* 26, 1651–1686. doi: 10.1214/aos/1024691352

Scornet, E., Biau, G., and Vert, J.-P. (2015). Consistency of random forests. *Ann. Stat.* 43, 1716–1741. doi: 10.1214/15-AOS1321

Seebens, H., Blackburn, T. M., Dyer, E. E., Genovesi, P., Hulme, P. E., Jeschke, J. M., et al. (2017). No saturation in the accumulation of alien species worldwide. *Nat. Commun.* 8:14435. doi: 10.1038/ncomms14435

Sfikas, G. (1987). *Wild Flowers of Crete*. Athens: Efstathiadis Group.

Simberloff, D., Martin, J. L., Genovesi, P., Maris, V., Wardle, D. A., Aronson, J., et al. (2013). Impacts of biological invasions: what's what and the way forward. *Trends Ecol. Evol.* 28, 58–66. doi: 10.1016/j.tree.2012.07.013

Strobl, C., and Boulesteix, A. (2007). Bias in random forest variable importance measures: illustrations, sources and a solution. *BMC Bioinformatics* 8:25. doi: 10.1186/1471-2105-8-25

Thuiller, W., Richardson, D. M., Rouget, M., Procheş, Ş., Wilson, J. R. U., Procheş, S., et al. (2006). Interactions between environment, species traits, and human uses describe patterns of plant invasions. *Ecology* 87, 1755–1769. doi: 10.1890/0012-9658(2006)87[17552.0.CO;2]

Tittensor, D. P., Walpole, M., Hill, S. L. L., Boyce, D. G., Britten, G. L., Burgess, N. D., et al. (2014). A mid-term analysis of progress toward international biodiversity targets. *Science* 346, 241–244. doi: 10.1126/science.1257484

Trombetti, M., Katsanevakis, S., Deriu, I., and Cardoso, A. C. (2013). EASIN-Lit: a geo-database of published alien species records. *Manag. Biol. Invasions* 4, 261–264. doi: 10.3391/mbi.2013.4.3.08

Trzaska, S., and Schnarr, E. (2014). *A Review of Downscaling Methods for Climate Change Projections*. United States Agency Int. Dev.

Tsanis, I. K., Koutroulis, A. G., Daliakopoulos, I. N., and Jacob, D. (2011). Severe climate-induced water shortage and extremes in Crete. *Clim. Change* 106, 667–677. doi: 10.1007/s10584-011-0048-2

Turland, N., Chilton, L., and Press, J. (1993). *Flora of the Cretan Area: Annotated Checklist and Atlas*. London: JR Press.

Vezza, P., Muñoz-Mas, R., Martinez-Capel, F., and Mouton, A. (2015). Random forests to evaluate biotic interactions in fish distribution models. *Environ. Model. Softw.* 67, 173–183. doi: 10.1016/j.envsoft.2015.01.005

Vilà, M., Espinar, J. L., Hejda, M., Hulme, P. E., Jarošík, V., Maron, J. L., et al. (2011). Ecological impacts of invasive alien plants: a meta-analysis of their effects on species, communities and ecosystems. *Ecol. Lett.* 14, 702–708. doi: 10.1111/j.1461-0248.2011.01628.x

Vitale, M., Proietti, C., Cionni, I., Fischer, R., and De Marco, A. (2014). Random forests analysis: a useful tool for defining the relative importance of environmental conditions on crown defoliation. *Water Air Soil Pollut.* 225:1992. doi: 10.1007/s11270-014-1992-z

Vogiatzakis, I. N., Griffiths, G. H., and Mannion, A. M. (2003). Environmental factors and vegetation composition, Lefka Ori massif, Crete, S. Aegean. *Glob. Ecol. Biogeogr.* 12, 131–146. doi: 10.1046/j.1466-822X.2003.00021.x

Wager, S., Hastie, T., and Efron, B. (2014). Confidence intervals for random forests: the jackknife and the infinitesimal jackknife. *J. Mach. Learn. Res.* 15, 1625–1651. Available online at: http://jmlr.org/papers/v15/wager14a.html

Wang, X. H., Liu, Y. J., and Guo, L. N. (2014). The invasive species risk assessment and prediction system based on GIS. *Appl. Mech. Mater*. 675–677, 1052–1057. doi: 10.4028/www.scientific.net/AMM.675-677.1052

Conflict of Interest Statement: The author ID was employed by company TM Solutions.

The other authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest

frontiers in ICT

TECHNOLOGY REPORT
published: 27 July 2017
doi: 10.3389/fict.2017.00020

Handling Big Data of Alien Species in Europe: The European Alien Species Information Network Geodatabase

Ivan Deriu, Fabio D'Amico, Konstantinos Tsiamis, Eugenio Gervasini and Ana Cristina Cardoso*

European Commission, Joint Research Centre – Directorate for Sustainable Resources, Water & Marine Resources Unit, Ispra, Italy

Building and managing large datasets of alien species is crucial to research, management, and control of biological invasions. To this end, the European Alien Species Information Network (EASIN) platform aggregates, integrates, and harmonizes spatio-temporal data regarding alien species in Europe, including both invasive and non-invasive alien species. These data are stored in the EASIN Geodatabase after their harvesting from relevant sources in the frame of a global and European databases partnership and scientific literature. The ownership of the data remains with its source, which is properly cited and linked. The process of data harvesting is performed through the EASIN Data Broker system, which retrieves the information related to alien species data in Europe and stores them in a normalized database structure. Data are subsequently refined through validation, cleansing, and standardization processes and finally stored in the EASIN Geodatabase. All data are finally visualized and shown in occurrence maps at different levels of spatial visualization. Analysis of the data contained in the EASIN Geodatabase through flexible web services offered by the system has already provided useful input in scientific works and policies on biological invasions. Data from European Union (EU) member state official surveillance systems, within the framework of the EU Regulation 1143/2014 on invasive alien species, are expected to contribute to the update of the EASIN Geodatabase. In addition, data from citizen science initiatives will further enrich the Geodatabase after appropriate validation. In this article, we describe and discuss the technical aspects, data flow and capabilities of the EASIN Geodatabase.

Keywords: alien, dataset, European Alien Species Information Network, Europe, Geodatabase, network

OPEN ACCESS

Edited by:
Aristides (Aris) Moustakas,
Queen Mary University of London,
United Kingdom

Reviewed by:
Tim Adriaens,
Research Institute for Nature and Forest, Belgium
Panayiotis Dimitrakopoulos,
University of the Aegean, Greece

***Correspondence:**
Ivan Deriu
ivan.deriu@ext.jrc.ec.europa.eu

Specialty section:
This article was submitted to Environmental Informatics, a section of the journal Frontiers in ICT

Received: *30 March 2017*
Accepted: *05 July 2017*
Published: *27 July 2017*

Citation:
Deriu I, D'Amico F, Tsiamis K, Gervasini E and Cardoso AC (2017) Handling Big Data of Alien Species in Europe: The European Alien Species Information Network Geodatabase. Front. ICT 4:20. doi: 10.3389/fict.2017.00020

INTRODUCTION

There are more than 14,000 alien species that have been reported so far in Europe (Katsanevakis et al., 2015). About 10% of them correspond to invasive alien species (IAS) based on the Tens Rule (Williamson and Fitter, 1996). IAS threaten or adversely impact upon biodiversity and related ecosystem services of Europe [Millennium Ecosystem Assessment (MEA), 2005; Ricciardi et al., 2013; EU, 2014; Jeschke et al., 2014]. As a conservative estimate, IAS cost the European Union (EU) Member States €12 billion in damages on annual basis (Kettunen et al., 2009) but cumulated costs probably reach €20 billion per year (IEEP, 2016). In addition, there is an increasing trend toward introduction of new alien species (Essl et al., 2015; Roques et al., 2016; Seebens et al., 2017). Recognizing the need for a coordinated set of actions to prevent, control, and mitigate the adverse impact of IAS on biodiversity and related ecosystem services, the European Parliament and the Council have adopted the EU Regulation no. 1143/2014 (EU, 2014; hereafter referred to as the IAS Regulation) on the prevention and management of the introduction and spread of IAS, which entered into force on January 1, 2015.

Managing large datasets of alien species is crucial to research, management, and control of biological invasions. Handling such large datasets poses a great challenge, since the information available is generally scattered across many different information systems and databases (Panov and Gollasch, 2004; Gatto et al., 2013). In addition, data format and information about the data (metadata) are frequent limitations to the interoperability of data repositories and information systems (Katsanevakis et al., 2012).

The European Alien Species Information Network (EASIN[1]) aims to provide a single repository of alien species data for accessing all the information necessary to underpin alien-species-related policy and evidence-based decision-making (Katsanevakis et al., 2013a). A key component of the EASIN system is the EASIN Catalog: a comprehensive list of alien species in Europe, currently including information on more than 14,000 taxa in a wide range of environments. Both invasive and non-invasive alien species are included. IAS are tagged as "high-impact" species in EASIN and are considered those that are highlighted by the lists of Delivering Alien Invasive Species Inventories for Europe, Global Invasive Species Database, NOBANIS (European Network on IAS), CABI's Invasive Species Compendium, MedPAN (Network of Marine Protected Areas managers in the Mediterranean), and SEBI-2010 (list of worst IAS threatening biodiversity in Europe; Streamlining European 2010 Biodiversity Indicators), to cause severe ecological and socio-economic impacts in Europe and/or globally. Other species are tagged as "low/unknown impact" species (see also Katsanevakis et al., 2012). In the latest version of EASIN (v.4.1), there are 908 species tagged as "high-impact," while 13,102 species are tagged as "low/unknown impact."

For each alien species of the EASIN Catalog information is available on the year and country of the first record in Europe, alien status (alien, cryptogenic, questionable), native range, taxonomy, synonyms, common names, environment, pathways, vectors, and impact. Links to factsheets are also provided for selected taxa (Katsanevakis et al., 2015). Given the vast amount of information recorded in EASIN and the need for constant updating and revision, an Editorial Board has been established to ensure the quality and updates of the information contained in the EASIN Catalog (Tsiamis et al., 2016).

Besides the EASIN Catalog, EASIN provides aggregated and integrated spatial data on alien species in Europe as well as online mapping tools for the retrieval of these data, through the EASIN Geodatabase. All spatial data are visualized and shown in occurrences maps at different levels (country, grid 10 km × 10 km, river basin, marine ecoregion). In the current paper, we demonstrate the approach, functionality, and technical background of the EASIN Geodatabase, which is the backbone of the EASIN network, responsible for the collection, quality control, harmonization, integration, storage, visualization, and dissemination of large datasets on alien species in Europe, coming from various sources.

DESCRIPTION OF THE EASIN GEODATABASE

Data Sources

The EASIN Geodatabase hosts about 50 millions of alien species spatial data (i.e., records) across Europe, all originating from a network of data sources, named as EASIN *Data Partners* (**Table 1**). These are global, regional, and national databases or repositories that have agreed to provide EASIN with spatial information of alien species in Europe. These sources do not always contain alien species data exclusively, but they might host broader species biodiversity information, including also native species in Europe, such as in the case of Global Biodiversity Information Facility. However, EASIN performs a targeted harvesting only to species which are considered alien or partly alien in the whole European scale (see Data Harvesting). Ownership of the data remains with its source, which is properly cited and linked in the EASIN Geodatabase. The EASIN Data Partners gain increased visibility and networking possibilities through EASIN, and can also benefit from mutual data exchange. Additional and updated spatial data are being gathered through EASIN-Lit (Trombetti et al., 2013), which contributes to enriching the Geodatabase with geo-referenced data published in the scientific literature. By

TABLE 1 | European Alien Species Information Network (EASIN) Data Partners and details on their technology, access, and update schedule in the EASIN Geodatabase.

Data partner	Technology	Access	Update schedule
Global Biodiversity Information Facility	Web service	Online	Monthly
Global Invasive Species Information Network	MS access DB	Offline	No schedule
Regional Euro-Asian Biological Invasions Center	MS SQL DB	Online restricted	Monthly
Hellenic Network on Aquatic Invasive Species (ELNAIS)	ArcGIS DB	Offline	No schedule
International Commission for Scientific Exploration of the Mediterranean Sea (CIESM)	ArcGIS DB	Offline	No schedule
EASIN-LIT	Excel file/ ArcGIS DB	Offline	Monthly
HCMR-European Environmental Agency	Excel file	Offline	No schedule
International Union for Conservation of Nature online information system for monitoring invasive non-native species in marine protected areas (MEDMIS)	PostgreSQL DB	Online restricted	Monthly
Marine Mediterranean Invasive Alien Species	Excel file	Offline	No schedule
Norwegian Biodiversity Information Center	Excel file	Offline	No schedule

---

[1] http://easin.jrc.ec.europa.eu.

March 2017, 11 data sources have been included as EASIN Data Partners.

Data Harvesting

The targeted information for the EASIN Geodatabase is the alien species spatial data in Europe, regardless if they are invasive or not. These can also include partly native species in Europe, i.e., species which are alien in some areas of Europe but native to some others. The process of harvesting information is done through the EASIN Data Broker system (**Figure 1**). This system is able to retrieve heterogeneous species data from different categories of data sources and formats and store them in a normalized database structure. In more detail, it screens and harvests information from the EASIN Data Partners based on the scientific names of the alien species listed in the EASIN Catalog. The EASIN Data Broker system identifies only the binomial of each species, without taking into consideration the species scientific authorities. As a result, typos and/or various formats of the species authorities in other sources do not affect the harvesting of the targeted information. In addition, scientific synonyms of each species are considered during harvesting. This is achieved through the information contained in the EASIN Catalog. For each species in the Catalog all scientific synonyms are also included, based on the Integrated Taxonomic Information System for terrestrial and freshwater species, and the World Register of Marine Species (WoRMS) for marine species. Thus, the EASIN Data Broker system can spot and harvest information based not only on the species binomial but also on its scientific synonyms.

Once the EASIN Data Broker system spots the binomial of a species within a dataset of an EASIN Data Partner, it will harvest targeted information about the specific species: (a) its location (name of the area and/or coordinates), (b) its date of collection (year), and (c) a reference to the original source of the record. Inside the EASIN system each Data Partner has a specific database and unique identifier. For each new data harvesting of a Data Partner a new dataset with related metadata is created.

Data Processing

Once harvested, the data are processed through different transformation steps: validation, cleansing, standardization, and geocoding (**Figure 1**). The validation and cleansing processes verify if the harvested data contain all the required information

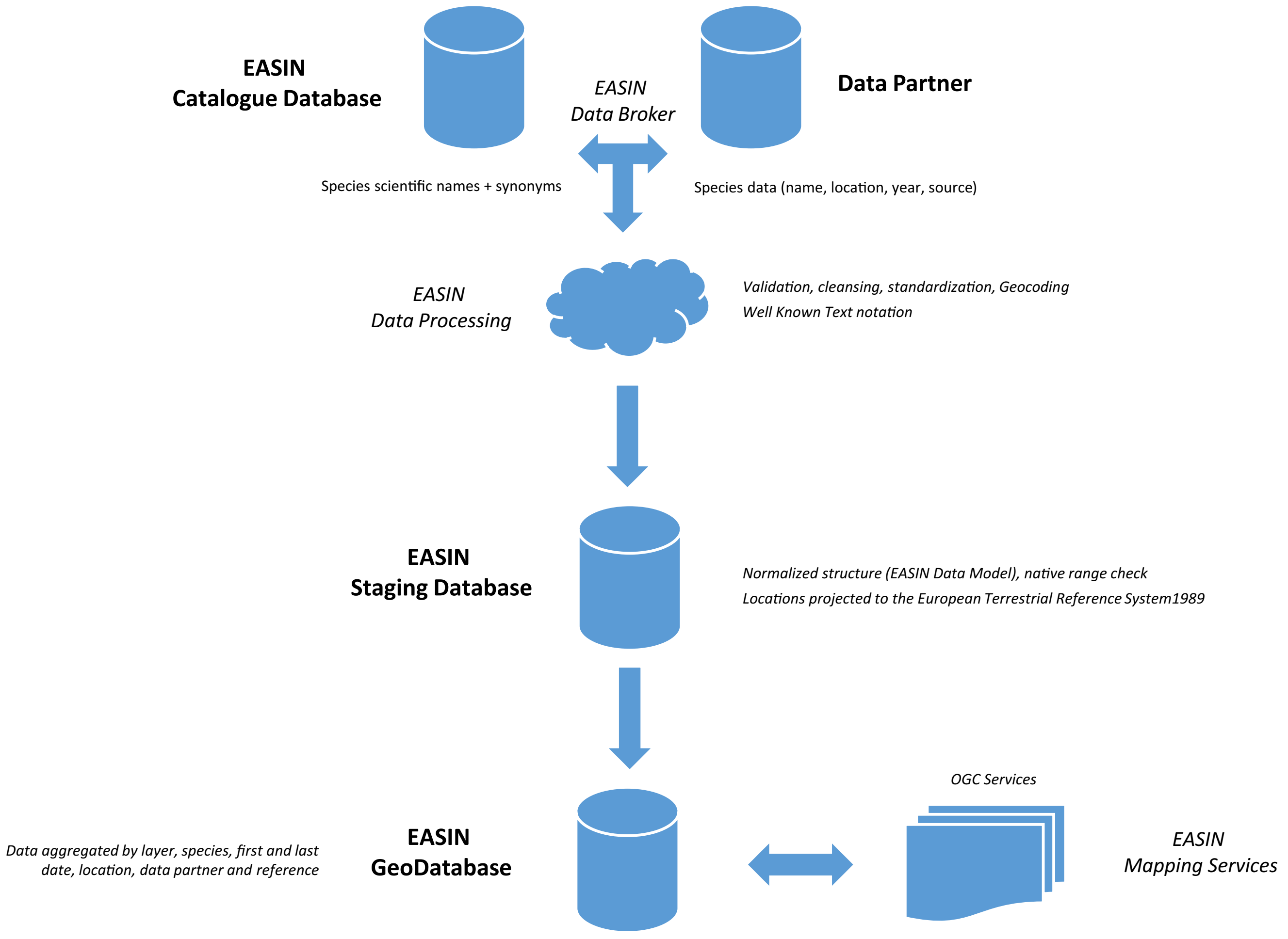


FIGURE 1 | Data flow in the European Alien Species Information Network (EASIN) Geodatabase.

(species name, date, location, reference), and if the reported location falls in one of the countries included in the EASIN geographical coverage of Europe.[2] In case the harvested data contains spatial information through coordinates, a further verification checks if the reported location corresponds with the relevant environment of the species. This is achieved through the information contained in the EASIN Catalog for the environment of each species (terrestrial, freshwater, marine, oligohaline—see also Katsanevakis et al., 2015). As a result, the verification check can spot a mismatch of a species environment compared with a geo-reference occurrence. Thus, a record containing a terrestrial or freshwater species reported in the sea will be discarded. Similarly, a marine species reported in a terrestrial or freshwater location will be also excluded. After validation and cleansing, the information is standardized following the EASIN Data model. This model organizes and normalizes the information coming from the Data Partners. During this process unique identifiers for locations and references are created.

The format of the aggregated spatial information regarding the locations is varying across the EASIN Data Partners. To this end, a conversion of this information is performed through the Well Known Text (WKT[3]) markup language, ensuring interoperability of the spatial information aggregated. Spatial information retrieved from the Data Partners is stored in the coordinate system of World Geodetic System 1984 (WGS84).[4] In case the original information contains the location of a record in a different coordinate system, a re-projection process is performed through Arc GIS. When no coordinates are available, a conversion of cited place names or maps to a spatial format is performed with the best possible accuracy through the Google Geocoding.[5]

Staging Database

After the data have been processed, the information is moved to a normalized intermediate database, named EASIN Staging database (**Figure 1**). This structure allows transformation of the information in data formats that can be used for specific purposes, e.g., for serving the EASIN mapping (see Geodatabase and Mapping Services). In addition, through the Staging database the information can be transformed to the schemas required by the INSPIRE Regulation (INSPIRE Framework Directive 2007/2/EC; EU, 2007; INSPIRE, 2013). The EASIN Staging database contains the information about the location and the references in separate data collections in order to reduce data redundancy (e.g., one location may be reported in several different records) and improve data integrity (**Figure 2**). In this phase, the dates of the records are approximated to years, while locations are processed to contain the spatial information also in the European Terrestrial Reference System 1989 (ETRS89)[6] for coordinates. In the EASIN Staging database, the locations are intersected with the EASIN layers, in order to show in which country, grid cell 10 km × 10 km, marine ecoregion or river basin they fall. This process is fundamental for creating data aggregations by layers. The data are also processed against the information contained in the EASIN Catalog regarding the native range of a partly native species in Europe. This information is available in the EASIN Catalog in country-level. Once a partly native species is harvested from a Data Partner, the Staging database marks which occurrences fall inside countries which have been tagged as "native range" for the certain species in Europe. For example, if a partly native species in Europe is considered native in Italy, based on the information contained in the EASIN Catalog, then all harvested occurrences of the species falling in Italy will be set as native. The native range of the partly native species in Europe can be updated by the EASIN Editorial Board.

Geodatabase and Mapping Services

The EASIN Geodatabase is populated by extracting the data from the Staging database, aggregated by layer, species, first and last date, location, data partner, and reference (**Figure 1**). The result is a data format suitable for providing all the necessary information needed by the EASIN mapping services, which visualizes alien species occurrences' maps. The maps can be created at four levels, as many as the EASIN layers: by country, grid 10 km × 10km (based on EEA, 2012 reference grid), river basins, and marine ecoregions. When it comes to partly native species in Europe, their occurrences falling in countries which have been tagged as "native range" are depicted in a different color compared with the occurrences falling in countries tagged as "alien range." All the maps are exposed through Open Geospatial Consortium (OGC[7]) compliant services as Catalog Service for the Web, Web Map Service, and Web Feature Service.

Aiming at more tailored distribution maps of alien species in Europe, the EASIN Geodatabase can filter the related information based on the selection of the environment (terrestrial, freshwater, marine, oligohaline), species status (alien, cryptogenic, questionable), specific taxonomic groups (in various taxonomic levels from species to Kingdom), and pathways of introduction based on the scheme proposed by Hulme et al. (2008). Regarding the latter EASIN will gradually align its pathways categorization to the one proposed by the Convention of Biological Diversity (CBD, 2014) aiming at achieving synchronization and harmonizing of information on alien species pathways (Tsiamis et al., 2017a). Moreover, a distinction between high-impact and low-impact alien species is possible when depicting the occurrences maps.

IMPACT OF THE EASIN GEODATABASE

The EASIN Geodatabase has been used to support scientific research on biological invasions, addressing (a) distribution patterns (Katsanevakis et al., 2013b,c), (b) pathways and gateways of introduction (Katsanevakis et al., 2013c, 2014; Nunes et al., 2014, 2015; Essl et al., 2015; Roques, 2015; Pergl et al., 2017; Tsiamis et al., 2017a), (c) risk and impact assessments of invasive species (Katsanevakis et al., 2014, 2016; Perdikaris et al., 2016), and (d) Horizon Scanning exercises (Roy et al., 2015).

---

[2] https://easin.jrc.ec.europa.eu/Catalogue.

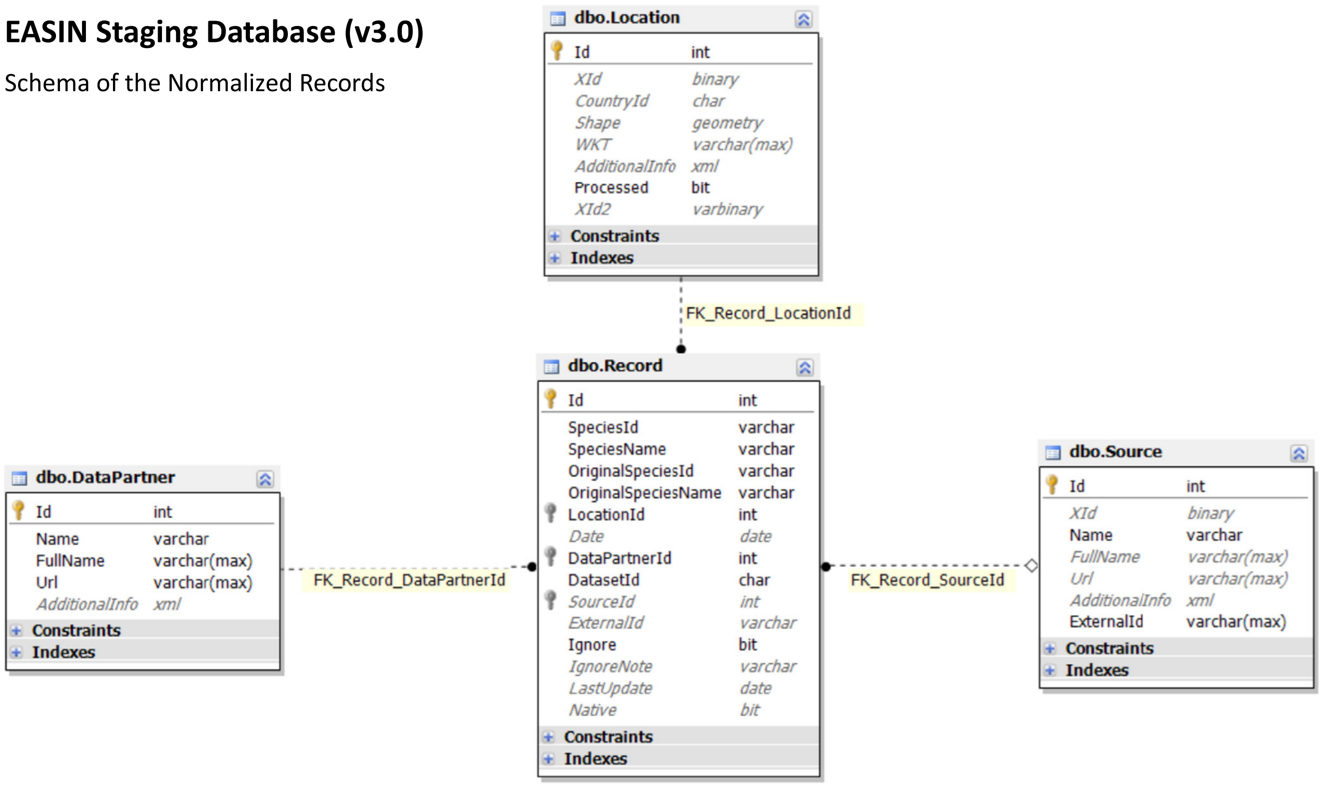
[3] https://en.wikipedia.org/wiki/Well-known_text.

[4] https://en.wikipedia.org/wiki/World_Geodetic_System.

[5] https://developers.google.com/maps/documentation/geocoding/intro.

[6] https://en.wikipedia.org/wiki/European_Terrestrial_Reference_System_1989.

[7] http://www.opengeospatial.org/.



dbo.Location

Column	Type
Id	int
XId	binary
CountryId	char
Shape	geometry
WKT	varchar(max)
AdditionalInfo	xml
Processed	bit
XId2	varbinary

dbo.Record

Column	Type
Id	int
SpeciesId	varchar
SpeciesName	varchar
OriginalSpeciesId	varchar
OriginalSpeciesName	varchar
LocationId	int
Date	date
DataPartnerId	int
DatasetId	char
SourceId	int
ExternalId	varchar
Ignore	bit
IgnoreNote	varchar
LastUpdate	date
Native	bit

dbo.DataPartner

Column	Type
Id	int
Name	varchar
FullName	varchar(max)
Url	varchar(max)
AdditionalInfo	xml

dbo.Source

Column	Type
Id	int
XId	binary
Name	varchar
FullName	varchar(max)
Url	varchar(max)
AdditionalInfo	xml
ExternalId	varchar(max)

FIGURE 2 | Schema of the normalized records in the European Alien Species Information Network (EASIN) Staging Database, containing the information about the location and the references in separate data collections.

European Alien Species Information Network constitutes the core information system supporting MS in the implementation of the IAS Regulation (EU, 2014, Art. 25). The 37 IAS of Union concern, published in the Commission Implementing Reg. 1141/2016 in application of the IAS Regulation, are all included in EASIN and tagged distinctively as "Union concern." The information stored in the EASIN Geodatabase regarding these 37 taxa has been used as the basis for the establishment of a distribution baseline of those species in terms of the IAS Regulation. This baseline is an important tool supporting the implementation of the IAS Regulation, the Art. 16 in particular regarding early detections of the IAS of Union concern, and also provides a factual basis for the review of the application of the IAS Regulation (Tsiamis et al., 2017b). In addition, information found in the EASIN Geodatabase has been used for the review of the EU Biodiversity Strategy, especially when it comes to spatial patterns of the main pathways of introduction of alien species into Europe, including both invasive and non-invasive species (EU, 2015).

CHALLENGES AND SOLUTIONS

Handling heterogeneous information coming from different sources and providing maps to the users requires to address two main challenges: (a) harmonizing of the information coming from different data sources and (b) minimizing the time required for showing a map to the user.

Normalizing the data (references and locations) and using a unique format for storing the information has been crucial for further managing the data and producing integrated maps and reports. In particular, aside the effort on implementing the data processes, the solution has been the choice to use only one projection (WGS84) for uniquely identifying the locations. Each EASIN Data Partner has its own particular data model, storage type, publishing approach, and standards that require a specific EASIN data brokering procedure (**Table 1**).

The EASIN Geodatabase output is to provide to the users occurrences maps of alien species in Europe. Taking into account the vast information included in the Geodatabase and the processing of it, there has been a challenge to visualize maps in the EASIN interface through the shortest possible time. In order to improve the reading performance, the first solution implemented has been the de-normalizing and grouping of some information. After the massive update of EASIN Geodatabase during 2016, which reached the amount of almost 50 million of data processed, it has been clear that the de-normalization solution was not enough anymore, considering also the large amount of space occupied by the redundant information. The solution has been found in the creation of a "dynamic query pre-cooking." Every time a user sends a query to the Geodatabase, the result of the query is stored in a specific location. Thus, a unique identifier is built for each query result through a hash algorithm. If a user sends a query already present in the EASIN Geodatabase the system retrieves the final result without further processing. Any new release of the Geodatabase results in deleting the current pre-cooked queries, as the aggregations in the new release may differ from the previous one.

Maps provided by the EASIN Geodatabase are exclusively based on the information provided by the EASIN Data Partners. However, this information might be biased due to varying monitoring efforts, geographical coverage, environment, and taxonomic groups addressed in each original dataset of an EASIN Data Partner. As a principle, EASIN targets to be enlarged with as many Data Partners as possible hosting information on alien species in Europe. Then, it is up to the user to select only the Data Partners which are considered more appropriate for an optimal visualization of occurrences of alien species in Europe and correct the data from potential biases.

FUTURE DEVELOPMENTS

Up to March 2017 there have been 11 EASIN Data Partners, but at least 7 new ones will be integrated in EASIN by the end of 2017: Icelandic Institute of Natural History, Academy of Science of Moldova, National Biodiversity Data Center—Ireland, Great Britain Non-Native Species Secretariat, Croatian Agency for the Environment and Nature, Azorean Biodiversity Portal, Life Project STOPVESPA.

The EASIN Geodatabase will be further enriched by records of IAS of Union concern [EU, 2014, Art. 3(3) and 4(1)] submitted by the EU Member States' Competent Authorities for the IAS Regulation in an Early Warning and Rapid Eradication Notification System (NOTSYS), developed with close links to EASIN. Through this system, the European Commission and all Member States are notified about new occurrences of IAS of Union concern detected on EU territory. NOTSYS is also used for reporting eradication measures as well as their effectiveness, as foreseen in Art. 16-17 of the IAS Regulation.

In addition, it is under consideration that data of alien species coming from citizen science initiatives in Europe (e.g., the JRC smartphone application "*Invasive Aliens Species in Europe*"—Tsiamis et al., 2017c), which are increasingly recognized as an important source of alien species data (Thiel et al., 2014; Adriaens et al., 2015; Daume, 2016), will further enrich the EASIN Geodatabase after appropriate validation. In this context, the EASIN Geodatabase could offer an aggregation and harmonization point for data coming from all citizen science projects related with alien species in Europe.

Prioritization of areas which are most sensitive and/or susceptible to biological invasions is critical for their effective management (McGeoch et al., 2015). To this end, the EASIN Geodatabase will include an additional layer within its mapping services on the Natura 2000 conservation network, in the frame of the Birds Directive (EU, 2009) and the Habitats Directive (EU, 1992). Moreover, the inclusion of the Marine Strategy Framework Directive (EU, 2008) marine regions as a distinct layer in EASIN maps is also under consideration. This would facilitate suitable filtering of the related information regarding marine alien species in terms of the Directive's purposes. In addition, the linkage with the IAS Regulation will not be limited exclusively to the 37 IAS of Union concern, but to a wider context of the EASIN Geodatabase species, taking into account the possible inclusion of additional species in the Union concern list, and the establishment of species of "Regional," "Member State" concern, and IAS of EU "outermost regions" (Art. 11, 12, and 6, respectively, of the IAS Regulation).

Speeding up the data processing is a critical aspect for any procedure that integrates data from different sources and requires constant update. In particular, when spatial information is shown on aggregated maps, as in the case of EASIN, a less time-demanding data process is required. In order to retrieve huge amounts of data, such as in the case of the grid 10 km × 10 km, the usual approach of processing the information record by record cannot be the most efficient approach. A location is usually shared among different records, thus the number of locations is less than the number of records. Therefore, locations can be processed faster than records. To this end, the locations must be normalized, as in the EASIN Staging database, before being processed. By adopting this approach in EASIN, the time required for processing millions of records can be reduced from weeks to hours.

CONCLUSION

Tackling biological invasions effectively relies on the availability of up-to-date scientific information, and requires the sharing of knowledge on the topic in order to enhance science-based decision-making (Panov et al., 2011; Katsanevakis et al., 2012; McGeoch et al., 2012; Roy et al., 2014; Groom et al., 2017a). In addition, it is crucial to have a single aggregation point, where available knowledge on alien species from various data sources is integrated, harmonized, and standardized (Panov and Gollasch, 2004; Katsanevakis et al., 2013a; Ojaveer et al., 2014). This knowledge should be openly displayed in ways that is findable, accessible and interoperable, encouraging exchange, and dissemination of data with other existing data platforms (Groom et al., 2015, 2017b; Lucy et al., 2016), in line with the EU Open Science Strategic Priority, and the plan of the Commission to develop the European Open Science Cloud (EU, 2016).

European Alien Species Information Network is the core of the information support system supporting the implementation of the IAS Regulation, while at the same time there has been a linkage with the EU Biodiversity policy. Connection with other policies on alien species, such as the MSFD and Natura 2000 sites is also under consideration. Therefore, it is crucial that the information provided by the EASIN Geodatabase is of high quality, updated, accessible, and easily processed by all interested stakeholders.

European Alien Species Information Network, through the EASIN Geodatabase, offers both a single aggregation point of alien species spatial data and flexible services for mapping alien species occurrences within Europe. EASIN works in partnership with European and global data providers to facilitate access to key data and information on alien species, promotes the principles of open-source, ensures accreditation of data publishers and data owners, highlights the work of other initiatives, increases their visibility, and facilitates direct access to the original information/data.

The number of alien species data processed by the EASIN Geodatabase will considerably increase in the near future due to the incorporation of new EASIN Data Partners, NOTSYS data, and citizen science data. This additional amount of information will test the efficiency and endurance of the EASIN Geodatabase, and possibly new adaptations and technical solutions might be needed. In addition, the continuous growth of the EASIN Geodatabase will test the soundness of the statistical correlations performed based on EASIN data, such as in the case of the analyses on pathways of introduction or distribution patterns of alien species. This is because bigger data sets could result in higher bias and bring up spurious correlations (Silver, 2012; Donoho and Jin, 2015) with possible implications to outcomes crucial to biological invasions research and to related policies. To this end, careful selection of data and appropriate statistical design should be ensured in order to limit correlated errors when handling big data sets (see also Moustakas, 2017 and references therein).

AUTHOR CONTRIBUTIONS

All authors have contributed to the present manuscript. ID and FD focused in the design of the EASIN Geodatabase, AC in the concept of the work, and KT and EG in drafting the manuscript.

ACKNOWLEDGMENTS

The authors wish to thank the European Commission's Directorate-General for Environment (DG ENV) for their support.

REFERENCES

Adriaens, T., Sutton-Croft, M., Owen, K., Brosens, D., van Valkenburg, J., Kilbey, D., et al. (2015). Trying to engage the crowd in recording invasive alien species in Europe: experiences from two smartphone applications in northwest Europe. *Manag. Biol. Invasion* 6, 215–225. doi:10.3391/mbi.2015.6.2.12

CBD. (2014). *Pathways of Introduction of Invasive Species, Their Prioritization and Management*, UNEP/CBD/SBSTTA/18/9/Add.1.Vol. 6. Montreal, Canada: UNEP/CBD/SBSTTA. 1–18.

Daume, S. (2016). Mining Twitter to monitor invasive alien species—an analytical framework and sample information topologies. *Ecol. Inform.* 31, 70–82. doi:10.1016/j.ecoinf.2015.11.014

Donoho, D., and Jin, J. (2015). Higher criticism for large-scale inference, especially for rare and weak effects. *Stat. Sci.* 30, 1–25. doi:10.1214/14-STS506

EEA. (2012). *European Environmental Agency Reference Grid*. Available at: http://www.eea.europa.eu/data-and-maps/data/eea-reference-grids-1

Essl, F., Bacher, S., Blackburn, T., Booy, O., Brundu, G., Brunel, S., et al. (2015). Crossing frontiers in tackling pathways of biological invasions. *Bioscience* 65, 769–782. doi:10.1093/biosci/biv082

EU. (1992). Council directive 92/43/EEC of 21 May 1992 on the conservation of natural habitats and of wild fauna and flora. *Off. J. Eur. Union* L206, 7–50.

EU. (2007). Directive 2007/2/EC of the European parliament and of the council of 14 March 2007 establishing an infrastructure for spatial information in the European community (INSPIRE). *Off. J. Eur. Union* L108, 1–14.

EU. (2008). Directive 2008/56/EC of the European parliament and of the council of 17 June 2008 establishing a framework for community action in the field of marine environmental policy (marine strategy framework directive). *Off. J. Eur. Union* L164, 19–40.

EU. (2009). Directive 2009/147/EC of the European parliament and of the council of 30 November 2009 on the conservation of wild birds. *Off. J. Eur. Union* L184, 7–25.

EU. (2014). Regulation (EU) No 1143/2014 of the European parliament and of the council on the prevention and management of the introduction and spread of invasive alien species. *Off. J. Eur. Union* L315, 35–55.

EU. (2015). Commission staff working document. EU assessment of progress in implementing the EU biodiversity strategy to 2020. Accompanying the document report from the commission to the European parliament and the council the mid-term review of the EU biodiversity strategy to 2020. *SWD* 187, 77.

EU. (2016). *Communication from the Commission to the European Parliament, the Council, the European Economic and Social Committee and the Committee of the Regions. European Cloud Initiative – Building a Competitive Data and Knowledge Economy in Europe*. COM/2016/0178 final. 13, Brussels: European Commission.

Gatto, F., Katsanevakis, S., Vandekerkhove, J., Zenetos, A., and Cardoso, A. C. (2013). Evaluation of online information sources on alien species in Europe – the need of harmonization and integration. *Environ. Manage.* 51, 1137–1146. doi:10.1007/s00267-013-0042-8

Groom, Q., Desmet, P., Vanderhoeven, S., and Adriaens, T. (2015). The importance of open data for invasive alien species research, policy and management. *Manag. Biol. Invasion* 6, 119–125. doi:10.3391/mbi.2015.6.2.02

Groom, Q., Weatherdon, L., and Geijzendorffer, I. R. (2017a). Is citizen science an open science in the case of biodiversity observations? *J. Appl. Ecol.* 54, 612–617. doi:10.1111/1365-2664.12767

Groom, Q., Adriaens, T., Desmet, P., Simpson, A., De Wever, A., Bazos, I., et al. (2017b). Seven recommendations to make your invasive alien species data more useful. *Front. Appl. Math. Stat.* 3:13. doi:10.3389/fams.2017.00013

Hulme, P. E., Bacher, S., Kenis, M., Klotz, S., Kühn, I., Minchin, D., et al. (2008). Grasping at the routes of biological invasions: a framework for integrating pathways into policy. *J. Appl. Ecol.* 45, 403–414. doi:10.1111/j.1365-2664.2007.01442.x

IEEP. (2016). *Institute for European Environmental Policy*. Available at: http://www.ieep.eu/work-areas/biodiversity/invasive-alien-species

INSPIRE. (2013). *D2.8.III.19 INSPIRE Data Specification on Species Distribution – Technical Guidelines*. INSPIRE Thematic Working Group Species Distribution.

Jeschke, J. M., Bacher, S., Blackburn, T. M., Dick, J. T. A., Essl, F., Evans, T., et al. (2014). Defining the impact of non-native species. *Conserv. Biol.* 28, 1188–1194. doi:10.1111/cobi.12299

Katsanevakis, S., Bogucarskis, K., Gatto, F., Vandekerkhove, J., Deriu, I., and Cardoso, A. C. (2012). Building the European alien species information network (EASIN): a novel approach for the exploration of distributed alien species data. *BioInvasions Rec.* 1, 235–245. doi:10.3391/bir.2012.1.4.01

Katsanevakis, S., Deriu, I., D'Amico, F., Nunes, A. L., Sanchez, S. P., Crocetta, F., et al. (2015). European alien species information network (EASIN): supporting European policies and scientific research. *Manag. Biol. Invasion* 6, 147–157. doi:10.3391/mbi.2015.6.2.05

Katsanevakis, S., Genovesi, P., Gaiji, S., Nyegaard Hvid, H., Roy, H., Nunes, A. L., et al. (2013a). Implementing the European policies for alien species – networking, science, and partnership in a complex environment. *Manag. Biol. Invasion* 4, 3–6. doi:10.3391/mbi.2013.4.1.02

Katsanevakis, S., Gatto, F., Zenetos, A., and Cardoso, A. C. (2013b). How many marine aliens in Europe? *Manag. Biol. Invasion* 4, 37–42. doi:10.3391/mbi.2013.4.1.05

Katsanevakis, S., Zenetos, A., Belchior, C., and Cardoso, A. C. (2013c). Invading European seas: assessing pathways of introduction of marine aliens. *Ocean Coast. Manage.* 76, 64–74. doi:10.1016/j.ocecoaman.2013.02.024

Katsanevakis, S., Tempera, F., and Teixeira, H. (2016). Mapping the impact of alien species on marine ecosystems: the Mediterranean Sea case study. *Diversity Distrib.* 22, 694–707. doi:10.1111/ddi.12429

Katsanevakis, S., Wallentinus, I., Zenetos, A., Leppäkoski, E., Çinar, M. E., Oztürk, B., et al. (2014). Impacts of marine invasive alien species on ecosystem services and biodiversity: a pan-European review. *Aquat. Invasions* 9, 391–423. doi:10.3391/ai.2014.9.4.01

Kettunen, M., Genovesi, P., Gollasch, S., Pagad, S., and Starfinger, U. (2009). *Technical Support to EU Strategy on Invasive Alien Species (IAS) – Assessment of the Impacts of IAS in Europe and the EU. Final Report for the European Commission*. Brussels, Belgium: Institute for European Environmental Policy (IEEP), 44.

Lucy, F. E., Roy, H., Simpson, A., Carlton, J. T., Hanson, J. M., Magellan, K., et al. (2016). INVASIVESNET towards an International Association for open knowledge on invasive alien species. *Manag. Biol. Invasion* 7, 131–139. doi:10.3391/mbi.2016.7.2.01

McGeoch, M. A., Genovesi, P., Bellingham, P. J., Costello, M. J., McGrannachan, C., and Sheppard, A. (2015). Prioritizing species, pathways, and sites to achieve conservation targets for biological invasion. *Biol. Invasions* 18, 299–314. doi:10.1007/s10530-015-1013-1

McGeoch, M. A., Spear, D., Kleynhans, E. J., and Marais, E. (2012). Uncertainty in invasive alien species listing. *Ecol. Appl.* 22, 959–971. doi:10.1890/11-1252.1

Millennium Ecosystem Assessment (MEA). (2005). *Ecosystems and Human Well-being: Biodiversity Synthesis*. Washington, DC, USA: World Resources Institute.

Moustakas, A. (2017). Spatio-temporal data mining in ecological and veterinary epidemiology. *Stoch. Environ. Res. Risk Assess* 31, 829–834. doi:10.1007/s00477-016-1374-8

Nunes, A. L., Katsanevakis, S., Zenetos, A., and Cardoso, A. C. (2014). Gateways to alien invasions in the European seas. *Aquat. Invasions* 9, 133–144. doi:10.3391/ai.2014.9.2.02

Nunes, A. L., Tricarico, E., Panov, V. E., Cardoso, A. C., and Katsanevakis, S. (2015). Pathways and gateways of freshwater invasions in Europe. *Aquat. Invasions* 10, 359–370. doi:10.3391/ai.2015.10.4.01

Ojaveer, H., Galil, B. S., Minchin, D., Olenin, S., Amorim, A., Canning-Clode, J., et al. (2014). Ten recommendations for advancing the assessment and management of non-indigenous species in marine ecosystems. *Mar. Policy* 44, 160–165. doi:10.1016/j.marpol.2013.08.019

Panov, V., and Gollasch, S. (2004). "Informational resources on aquatic alien species in Europe on the internet: present developments and future perspectives," in *UNESCO Proceedings of "The Colour of Ocean Data" Symposium*, IOC Workshop Report 188. (France, Paris).

Panov, V. E., Gollasch, S., and Lucy, F. (2011). Open-access journal *aquatic invasions*: an important part of the developing European information and early warning system on invasive alien species. *Aquat. Invasions* 6, 1–5. doi:10.3391/ai.2011.6.1.01

Perdikaris, C., Koustikos, N., Vardakas, L., Kommatas, D., Simonovic, P., Paschos, I., et al. (2016). Risk screening of non-native, translocated and traded aquarium freshwater fishes in Greece using Fish Invasiveness Screening Kit. *Fisheries Manag. Ecol.* 23, 32–43. doi:10.1111/fme.12149

Pergl, J., Pyšek, P., Bacher, S., Essl, F., Genovesi, P., Harrower, C. A., et al. (2017). Troubling travellers: are ecologically harmful alien species associated with particular introduction pathways? *NeoBiota* 32, 1–20. doi:10.3897/neobiota.32.10199

Ricciardi, A., Hoopes, M. F., Marchetti, M. P., and Lockwood, J. L. (2013). Progress towards understanding the ecological impacts of nonnative species. *Ecol. Monogr.* 83, 263–282. doi:10.1890/13-0183.1

Roques, A. (2015). Drivers and pathways of forest insect invasions in Europe, can we predict the next arrivals? *Accad. Naz. Ital. Entomol.* Anno LXIII, 145–150.

Roques, A., Auger-Rozenberg, M. A., Blackburn, T. M., Garnas, J. R., Pyšek, P., Rabitsch, W., et al. (2016). Temporal and interspecific variation in rates of spread for insect species invading Europe during the last 200 years. *Biol. Invasions* 18, 907–920. doi:10.1007/s10530-016-1080-y

Roy, H. E., Adriaens, T., Aldridge, D. C., Bacher, S., Bishop, J. D. D., Blackburn, T. M., et al. (2015). *Invasive Alien Species – Prioritising Prevention Efforts through Horizon Scanning*. ENV.B.2/ETU/2014/0016. Brussels: European Commission.

Roy, H. E., Preston, C. D., Harrower, C. A., Rorke, S. L., Noble, D., Sewell, J., et al. (2014). GB non-native species information portal: documenting the arrival of nonnative species in Britain. *Biol. Invasions* 16, 2495–2505. doi:10.1007/s10530-014-0687-0

Seebens, H., Blackburn, T. M., Dyer, E. E., Genovesi, P., Hulme, P. E., Jeschke, J. M., et al. (2017). No saturation in the accumulation of alien species worldwide. *Nat. Commun.* 8, 14435. doi:10.1038/ncomms14435

Silver, N. (2012). *The Signal and the Noise: Why So Many Predictions Fail-but Some Don't*. London: Penguin Books.

Thiel, M., Penna-Díaz, M. P., Luna-Jorquera, G., Salas, S., Sellanes, J., and Stotz, W. (2014). Citizen scientists and marine research: volunteer participants, their contributions, and projection for the future. *Oceanogr. Mar. Biol. Annu. Rev.* 52, 257–314. doi:10.1201/b17143-6

Trombetti, M., Katsanevakis, S., Deriu, I., and Cardoso, A. C. (2013). EASIN-Lit: a geo-database of published alien species records. *Manag. Biol. Invasion* 4, 261–264. doi:10.3391/mbi.2013.4.3.08

Tsiamis, K., Cardoso, A. C., and Gervasini, E. (2017a). The European alien species information network on the convention on biological diversity pathways categorization. *NEOBIOTA* 32, 21–29. doi:10.3897/neobiota.32.9429

Tsiamis, K., Gervasini, E., Deriu, I., D'Amico, F., Nunes, A., Addamo, A., et al. (2017b). *Baseline Distribution of Invasive Alien Species of Union Concern*. Ispra, Italy: Publications Office of the European Union. EUR 28596 EN.

Tsiamis, K., Gervasini, E., D'Amico, F., Deriu, I., Roglia, E., Shade, S., et al. (2017c). *Citizen Science Application, Invasive Alien Species in Europe*. Ispra, Italy: Publications Office of the European Union. EUR 28441 EN.

Tsiamis, K., Gervasini, E., D'Amico, F., Deriu, I., Katsanevakis, S., Crocetta, F., et al. (2016). The EASIN Editorial Board: quality assurance, exchange and sharing of alien species information in Europe. *Manag. Biol. Invasion* 7, 312–328. doi:10.3391/mbi.2016.7.4.02

Williamson, M., and Fitter, A. (1996). The varying success of invaders. *Ecology* 77, 1661–1666. doi:10.2307/2265769

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers
in Applied Mathematics and Statistics

ORIGINAL RESEARCH
published: 07 August 2017
doi: 10.3389/fams.2017.00016

A Space-Time Study of Hemorrhagic Fever with Renal Syndrome (HFRS) and Its Climatic Associations in Heilongjiang Province, China

Junyu He[1], **George Christakos**[1,2]*, **Wenyi Zhang**[3]* and **Yong Wang**[3]

[1] Institute of Islands and Coastal Ecosystems, Ocean College, Zhejiang University, Zhoushan, China, [2] Department of Geography, San Diego State University, San Diego, CA, United States, [3] Institute of Disease Control and Prevention, Academy of Military Medical Science, Beijing, China

Background: Hemorrhagic fever with renal syndrome (HFRS) is highly endemic in China, especially in Heilongjiang province (90% of all reported HFRS cases worldwide occur in China). The dynamic identification of high HFRS incidence spatiotemporal regions and the quantitative assessment of HFRS associations with climate change in Heilongjiang province can provide valuable guidance for HFRS monitoring, preventing and control. Yet, so far there exist very few and of limited scope quantitative studies of the spatiotemporal HFRS spread and its climatic associations in Heilongjiang province. Making up for this lack of quantitative studies is the reason for the development of the present work.

Method: To address this need, the well-known Bayesian maximum entropy (BME) method of space-time modeling and mapping together with its recently proposed variant, the projected BME (P-BME) method, were employed in this work to perform a composite space-time analysis and mapping of HFRS incidence in Heilongjiang province during the years 2005–2013. Also, using multivariate El Niño-Southern Oscillation index as a proxy, we proposed a combination of Hilbert-Huang transform and wavelet analysis to study the "HFRS incidence-climate change" associations.

Results: The main results of this work were two-fold: (1) three core areas were identified with high HFRS incidences that were spatially distributed and exhibited distinct biomodal temporal patterns in the eastern, western, and southern parts of Heilongjiang province; and (2) there exists a considerable association between HFRS incidence and climate change, particularly, an ∼6 months period coherency was clearly detected.

Conclusions: The combination of modern space-time modeling and mapping techniques (P-BME theory, Hilbert-Huang spectrum analysis, and wavelet analysis) used in this work led to valuable quantitative findings concerning the spatiotemporal spread of HFRS incidence in Heilongjiang province and its association with climate change. Our essential findings include the identification of three core areas with high HFRS incidences in Heilongjiang province, and considerable evidence that HFRS incidence is closely related to climate change.

Keywords: hemorrhagic fever with renal syndrome, spatiotemporal, mapping, Bayesian maximum entropy, Hilbert-Huang transformation, wavelet analysis

Edited by:
Aristides (Aris) Moustakas,
Queen Mary University of London,
United Kingdom

Reviewed by:
Petros Damos,
Aristotle University of Thessaloniki,
Greece
Alex Hansen,
Norwegian University of Science and
Technology, Norway

****Correspondence:***
George Christakos
gchristakos@zju.edu.cn
Wenyi Zhang
zwy0419@126.com

Specialty section:
This article was submitted to
Environmental Informatics,
a section of the journal
Frontiers in Applied Mathematics and
Statistics

Received: *27 March 2017*
Accepted: *18 July 2017*
Published: *07 August 2017*

Citation:
He J, Christakos G, Zhang W and Wang Y (2017) A Space-Time Study of Hemorrhagic Fever with Renal Syndrome (HFRS) and Its Climatic Associations in Heilongjiang Province, China. Front. Appl. Math. Stat. 3:16. doi: 10.3389/fams.2017.00016

INTRODUCTION

Hemorrhagic fever with renal syndrome (HFRS) is a rodent-borne zoonosis caused by Hantavirus (belonging to the Bunyaviridae family). In China, the Hantaan and Seoul viruses dominate HFRS infection, the leading rodent hosts of which are *Apodemus agrarius* and *Rattus norvegicus*, respectively [1, 2]. The virus is transmitted from rodents to humans via inhalation of aerosols contaminated by rodents' urine, saliva, excreta, and dung, possibly through ingestion of contaminated food and by direct contact of contaminated materials with broken skin or mucous membranes, or by rodent bites [3, 4]. Clinical HFRS manifestations include fever, headache, nausea, and abdominal pain. Complications, like adverse kidney effects and subsequent pulmonary edema, shock, renal insufficiency, encephalopathy, hemorrhages, and cardiac complications, can cause death [5, 6]. The disease goes through five stages: febrile, hypotensive shock, oliguric, polyuric, and convalescent, which last, respectively, 1–7 days, 1–3 days, 2–6 days, 2 weeks, and 3–6 months [7].

Historically, numerous HFRS-like cases have occurred in China going back to the tenth century AD. Currently, about 20,000–50,000 cases/year are reported in mainland China, which account for 90% of all reported cases worldwide [5, 8, 9]. The foci of HFRS often locates in rural areas, which constitute more than 70% of the total number of cases, because of poor housing conditions and abundant rodent hosts [10, 11]. According to the data of the National Health and Family Planning Commission of China, the HFRS death rate was 2.89% during 1950–2014 [7]. *Rattus norvegicus*, which hosts the Seoul virus, is regarded as one of the most damaging invasive species around the world, and it is closely associated with humans, particularly in large metropolitan areas [12, 13]. As HFRS remains a severe public health problem, it is necessary to study historical HFRS evidence to provide rigorous scientific support to current disease monitoring and control procedures. Yet, so far there exists a very limited number of quantitative studies regarding the spatiotemporal HFRS distribution and spread in Heilongjiang province. In fact, the reason for the development of the present work is to make up the lack of such quantitative studies.

Based on province-level data, a number of nationwide studies have been carried out in China. Their results indicated that the geographical distribution of HFRS incidences was clustered, particularly in the northeastern, central and eastern parts of China. The observed hotspots shifted and expanded from year to year, whereas most HFRS cases were mainly reported during the spring and the autumn-winter seasons [14–16]. In these studies, global indicators of spatial autocorrelation (GISA), local indicators of spatial association (LISA) and Kulldorff's scan statistic were employed to characterize the spatial variation of HFRS incidence during several time periods. However, these studies suffered from certain drawbacks: (a) they considered neither the temporal nor the combined space-time (spatiotemporal) correlation of HFRS incidences, and (b) a fine temporal resolution (i.e., monthly data) was assumed, whereas the spatial resolution used for mapping purposes was rather coarse (i.e., at the province-level).

For more than two decades, the Bayesian Maximum Entropy theory of space-time data analysis and mapping (BME, [17–19]) has been proven to provide efficient and cost-effective methods for characterizing, predicting and mapping disease attributes (such as disease incidence) in a composite space-time domain under conditions of *in-situ* uncertainty [20]. For example, Law et al. [21] used BME to qualitatively and quantitatively detect core areas with high syphilis incidence density in the city of Baltimore (USA) between the years 1994 and 2002; also, based on age-adjusted influenza mortality data at the county-level, Choi et al. [22] used BME to represent the space-time disease dependence structure of the disease, to map the influenza mortality rates and to assess the associated disease risk in the state of California (USA). In this work, we will use both the original BME and a BME variant (projected BME, P-BME) to study the spatiotemporal HFRS dependence pattern in Heilongjiang province, including the identification of particular disease features and the detection of high incidence areas.

Since HFRS is transmitted by reservoir hosts (especially rodents), it is expected that climatic factors (such as precipitation, temperature, humidity, and global climate pattern) should influence human HFRS morbidity by affecting the reproduction and abundance of rodents [3, 23–27]. To some extent, understanding climate change can offer a preliminarily assessment or an early warning concerning the epidemic situation. Few studies have investigated the intrinsic HFRS period and its inherent relationships with climate attributes and factors. In recent years, the Hilbert–Huang transformation (HHT, an adaptive method combining empirical mode decomposition and spectral analysis) has been developed for analyzing nonlinear and temporally non-stationary data, in general [28–30]. With this method, the intrinsic mode and intensity of a disease attribute is obtained that can provide useful insight regarding the temporal regulation of disease variation. Moreover, cross-wavelet transforms and wavelet coherence can explore the relationship between two series in the time-frequency domain [31]. As a matter of fact, the wavelet method has been employed in the past to study disease incidence and climate change. For example, Thai et al. [32] have found a strong non-stationary association between El Niño-Southern Oscillation (ENSO) indices and climate variables and the corresponding dengue incidence in the Binh Thuan province (Vietnam) during a 2–3 years period; also, Chowell et al. [33] have suggested that in Peru the dengue incidence is significantly linked with the seasonal cycle timing of mean temperature variations.

In view of the above considerations, the objective of this work is two-fold: (1) to investigate the characteristics and spatiotemporal distribution of HFRS incidence in Heilongjiang province (China) using the BME and P-BME methods; and (2) to assess the intrinsic mode and coherence similarities between the HFRS incidence spread and the time series of key climatic factors using the Hilbert-Huang spectrum method and wavelet analysis.

MATERIALS AND METHODS

Study Area and Data Collection

The Heilongjiang province is located in northeastern China, with an area of ∼473,000 Km^2 and a population of 38.35 million people. Remarkably, the Heilongjiang province is one of the highest HFRS morbidity regions in China [34]. The Heilongjiang basin includes four major river systems: the Heilong, Songhua, Wusuli, and Suifen rivers (**Figure 1**). Monthly data of HFRS cases (21,383 cases in total) were collected at 130 counties and districts during the January 2005–December 2013 period by the China Information System for Disease Control and Prevention (CISDCP). Demographic data for each county were obtained from the National Bureau of Statistics of China. Subsequently, the HFRS cases were population-standardized and used in the present work.

The multivariate ENSO Index (MEI) integrates 6 oceanic and meteorologic variables over the tropical Pacific region to represent global climatic cycles (i.e., El Niño-Southern Oscillation): sea-level pressure, zonal and meridional components of surface wind, sea surface temperature, surface air temperature, and total cloudiness fraction of sky [35]. The MEI has been used in the scientific literature to diagnose ENSO phenomena that can cause global climate variability, including world-wide correlations with temperature and precipitation data [36]. In the present study, the MEI is employed as a proxy for representing climatic factors in order to explore their effects on HFRS incidence (MEI data is available at https://www.esrl.noaa.gov/psd/enso/mei/table.html).

Spatiotemporal Analysis and Mapping

Methodologically, the standardized HFRS incidence was regarded as a spatiotemporal random field (S/TRF, [19, 37]), denoted as $X(\boldsymbol{p})$, with arguments $\boldsymbol{p} = (\boldsymbol{s}, t) \in R^2 \times T$ in a composite space-time domain, where $\boldsymbol{s} = (s_1, s_2) \in R^2$ denote the centroid coordinates of each administration unit and $t \in T$ denotes the time argument. In this quantitative modeling setting, space ($\boldsymbol{s}$) represents the order of co-existence and time (t) represents the order of successive existence of HFRS incidence distribution in Heilongjiang province. *In-situ* uncertainty manifests itself as an ensemble of possible HFRS realizations $\boldsymbol{x}$ regarding the space-time $X(\boldsymbol{p})$ distribution, where the likelihood that each one of these possible realizations occurs is expressed by the corresponding HFRS probability density function f. In the BME method, two main knowledge bases are considered: the general or core knowledge base (G-KB), and the site-specific knowledge base (S-KB). The G-KB includes theoretical models of the space-time HFRS mean and covariance (correlation), whereas the S-KB consists of hard (exact) and soft (uncertain) HFRS data [38]. The BME method uses the general knowledge available to generate the G–based (prior) probability density function, f_G, of HFRS incidence distribution. Subsequently, the S-KB is incorporated to generate the combined G- and S-based probability density function [19]

$$f_K(\boldsymbol{x}_k) = A^{-1} \int dS(\boldsymbol{x}_s) f_G(\boldsymbol{x}), \tag{1}$$

where $K = G+S$ denotes the total KB (core G and site-specific S), $\boldsymbol{x} = (\boldsymbol{x}_h, \boldsymbol{x}_s, \boldsymbol{x}_k)$ are HFRS realizations at the hard data points ($\boldsymbol{p}_h$), the soft data points ($\boldsymbol{p}_s$), and the unsampled (prediction) points ($\boldsymbol{p}_k$), and A is a normalization constant. After the probability density function f_K has been derived at all prediction points $\boldsymbol{p}_k$, various HFRS incidence, $X(\boldsymbol{p})$, estimates at these points are readily available, like the mean, the mode and the median $X(\boldsymbol{p})$ values at each $\boldsymbol{p}_k$.

It is widely-recognized that the joint spatial-temporal covariance of a disease attribute is rather difficult to calculate experimentally (primarily due to the limited number of sample points) and to fit to a theoretical covariance model [39, 40].

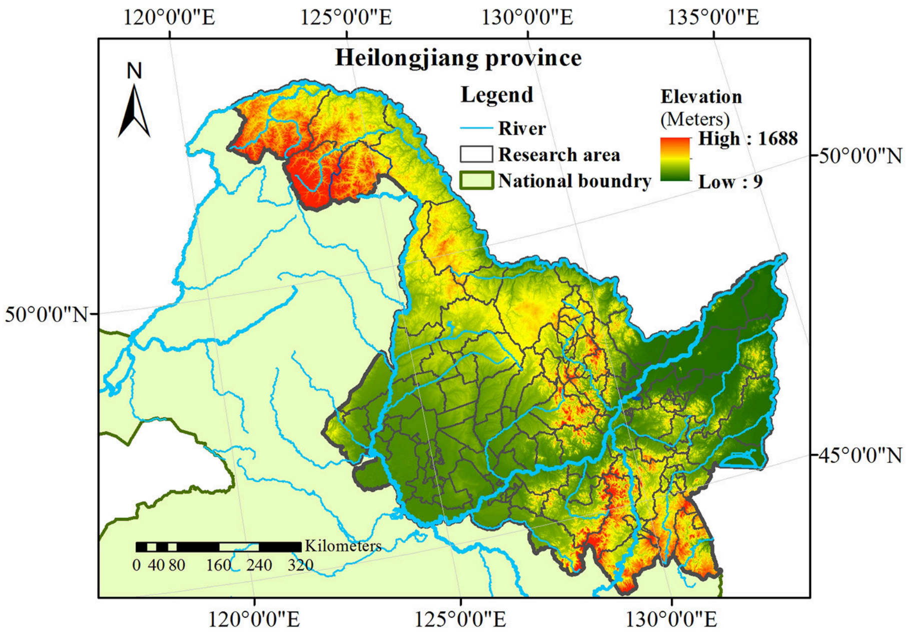


FIGURE 1 | Study area and delimiter of counties and districts.

Hence, it would be very useful to develop an improved method of space-time covariance fitting. Responding to this need, Christakos et al. [41] presented a BME variant, the projected BME (P-BME) method, which projects the disease incidence distribution (HFRS incidence in our case) from the original space-time disease domain $R^2 \times T$ onto a lower dimensionality traveling space domain R^2 by means of the simple set of equations

$$X(\boldsymbol{p}) = \hat{X}(\boldsymbol{s} - \boldsymbol{\upsilon}t, 0) = \hat{X}(\hat{\boldsymbol{s}}), \tag{2a}$$

$$c_X(\boldsymbol{h}, \tau) = c_{\hat{X}}(\boldsymbol{h} - \boldsymbol{\upsilon}\tau, 0) = c_{\hat{X}}(\hat{\boldsymbol{h}}), \tag{2b}$$

where $\hat{\boldsymbol{s}} = \boldsymbol{s} - \boldsymbol{\upsilon}t$ denotes the spatial coordinates of HFRS incidence values, $\hat{\boldsymbol{h}} = \boldsymbol{h} - \boldsymbol{\upsilon}\tau$ denotes the spatial lags (separation distances) between HFRS incidence values, and $\boldsymbol{\upsilon} = (\upsilon_1, \upsilon_2)$ is the HFRS traveling vector. Then, $\boldsymbol{p} = (s_1, s_2, t) \in R^2 \times T$ are the space-time coordinates in the original $R^2 \times T$ domain, which are matched one-to-one with the coordinates $\hat{\boldsymbol{s}} = (\hat{s}_1, \hat{s}_2)$ in the traveling R^2 domain according to the P-BME method, i.e., the following transformation of HFRS incidence domains is used,

$$(s_1, s_2, t) \in R^2 \times T \;\mapsto\; (\hat{s}_1, \hat{s}_2) \in R^2. \tag{2c}$$

Accordingly, $c_X(\boldsymbol{h}, \tau)$ and $c_{\hat{X}}(\hat{\boldsymbol{h}})$ are the HFRS incidence covariances in the space-time $R^2 \times T$ and the traveling (spatial) R^2 domains, respectively. The fact that these covariance functions are related by Equation (2b) provides a practical way to calculate $\boldsymbol{\upsilon}$ *in-situ*. Significant advantages of the P-BME method is that, after projection, the covariance $c_{\hat{X}}(\hat{\boldsymbol{h}})$ is located in the spatial (R^2) domain, where (a) it is easier to fit a theoretical HFRS covariance model to the data, (b) more choices of theoretical HRFS covariance models are allowed, and (c) HSRF incidence mapping is more accurate and computationally efficient than in the space-time ($R^2 \times T$) domain.

By way of a summary, the P-BME method combines the BME Equation (1) with the traveling Equations (2a–c) to derive HFRS incidence predictions across space-time. The mean absolute HFRS incidence prediction error can be used to evaluate the accuracy of model cross-validation prediction. For comparison purposes, the 10-fold cross-validation method will be used below to test the performance of the direct BME method and the P-BME method in the $R^2 \times T$ and the R^2 domains, respectively. More theoretical and technical details regarding BME and P-BME can be found in the cited literature.

Hilbert–Huang Transformation

The HHT can help discover certain characteristics of the cumulative HFRS incidence in the entire Heilongjiang province, together with the MEIs and the underlying rules of their variation. Basically, HHT consists of two steps: empirical mode decomposition (EMD), and Hilbert transformation (HT).

More specifically, EMD is used to extract the intrinsic mode functions (IMFs), each of which is independent of the others, from the raw series

$$Y(t) = \sum\nolimits_{m=1}^{n} c_m + r_n, \tag{3}$$

where c_m represents the IMF component, and r_n denotes residuals representing the raw series trend. We start by fitting the local minima and maxima values of the raw time series, $Y(t)$, using cubic splines, and the mean time series, $m_1(t)$, is defined (i.e., the spline mean). We also define the standard deviation (SD)

$$SD = \sum\nolimits_{t=0}^{T} \frac{|h_{1(k-1)}(t) - h_{1k}(t)|^2}{h^2_{1(k-1)}(t)}, \tag{4}$$

where $k = 1, 2, \ldots.$ denotes the number of times the process is repeated (by convention, $h_{10}(t) = Y(t)$). The first difference $h_{11}(t) = Y(t) - m_1(t)$ is considered as the first IMF if it satisfies the criterion that the SD is between 0.2 and 0.3 [28], otherwise, the process is repeated using $h_{11}(t)$ as a raw series until the criterion are met. After the first IMF is extracted, the difference between the first IMF and raw series is used to identify subsequent IMFs. Then, a monotonic series is defined as the residual of the raw series trend mentioned earlier.

The HT can be applied for each IMF above to obtain the analytical function and its polar form

$$\begin{aligned} z_i(t) &= c_i(t) + jH[c_i(t)], \\ z_i(t) &= a_i(t)\exp[j\theta_i(t)], \end{aligned} \tag{5}$$

respectively, where H denotes the Hilbert transform, and $a_i(t)$ and $\theta_i(t)$ are the IMF amplitude and phase, respectively. Each IMF is the real part of its corresponding analytical function $z_i(t)$. In view of the IMF summation, see Equation (3), the raw series can be calculated by

$$Y(t) = \mathrm{Re}\{\sum\nolimits_{i=1}^{n} a_i(t) e^{j\int \omega_i(t)dt}\}, \tag{6}$$

where $\omega_i(t)$ is the IMF instantaneous frequency. The Hilbert spectrum, $H(\omega, t)$, is subsequently defined as a frequency-time distribution with the amplitude of the raw series of Equation (6). And the marginal spectrum can be calculated by

$$h(\omega) = \int_0^T dtH(\omega, t). \tag{7}$$

Considering that EMD may cause mode mixing [42], in this work it was replaced by ensemble EMD (also denoted as EEMD). The process is as follows: add a white noise series into the raw series; use EMD to decompose the series with white noise into IMFs; repeat the two steps above (say, m times in total) with various white noise series; calculate the mean values of the corresponding m IMFs, which will be the final IMFs. The HHT can provide an amplified way of gaining insight into raw time series in the time-frequency domain, whereas the marginal spectrum offers an easy way to identify the cumulative amplitude distribution across different frequencies in a probabilistic sense [43]. In this work, the entire standardized HFRS incidences in Heilongjiang province and the MEI, 108 months in total, were analyzed by HHT for comparison. The software code used for this purpose can be downloaded from http://rcada.ncu.edu.tw/research1_clip_program.htm.

Wavelet Analysis

The coherency between standardized HFRS incidence and MEI expresses the association between these two variables, in which case the coherent time-frequency areas can be easily discovered in terms of wavelet analysis [31, 44].

Specifically, coherency can measure the covariation intensity between two series. For this purpose, the time series, say Y_1 and Y_2, are transformed by the continuous wavelet equation

$$W(\tau,\alpha) = \tfrac{1}{\sqrt{a}} \int_{-\infty}^{+\infty} Y(t)\, \psi^*(\tfrac{t-\tau}{a}) dt, \tag{8}$$

where $\psi(t)$ is the mother wavelet, and * denotes the complex conjugate. Subsequently, the two transformed series are cross-wavelet transformed by

$$W_{Y_1 Y_2} = W_{Y_1} W^*_{Y_2}. \tag{9}$$

Finally, the wavelet transform coherency across time-frequency space can be calculated by

$$R^2(\tau,\alpha) = \frac{|\alpha^{-1} W_{Y_1 Y_2}(\tau,\alpha)|^2}{S[\alpha^{-1}|W_{Y_1}(\tau,\alpha)|^2] S[\alpha^{-1}|W_{Y_2}(\tau,\alpha)|^2]}, \tag{10}$$

where α and τ are the scale factor and time shift, respectively, and S is a smoothing operator. As in the case of the BME method, more theoretical and technical details regarding HHT and wavelet analysis can be found in the cited scientific literature.

RESULTS

Spatiotemporal Analysis

A total of 14,040 space-time records were included in the spatiotemporal analysis of this work. In the $R^2 \times T$ domain, the HFRS incidence variability in Heilongjiang province was measured by means of the isotropic covariance plotted in **Figure 2A**. This space-time HFRS covariance combined two theoretical models, an exponential and a Gaussian (squared exponential) model, i.e.,

$$c_X(h,\, \tau) = e^{-\frac{h}{72\times 10^3} - (\frac{\tau}{2.6})^2}, \tag{11}$$

where 72×10^3 (m) and 2.6 ($months$) are, respectively, the spatial and temporal correlation ranges of the HFRS incidence distribution. The interpretation of the spatiotemporal covariance plots of **Figure 2A** implies that the distribution of HFRS cases during the period Jan 2005–Dec 2013 were controlled by spatial and temporal dependences. In quantitative terms, the covariance value [around 0.3 $(cases/10^5)^2$] at the time lag $\tau = 4$ ($months$) indicates a rather strong temporal dependence among the HFRS incidence values. On the other hand, the spatial neighborhood effect is about 200 (km), which also indicates a significant spatial dependence among incidence values.

By inserting Equation (11) into Equation (2b), the traveling coefficient was calculated to be $\upsilon = |\boldsymbol{\upsilon}| = -10650.89\tau$. Then, the HFRS incidence data points were projected from the $R^2 \times T$ domain onto the reduced dimensionality R^2 domain (Figure S1).

Following the projection process, the empirical HFRS covariance and the fitted covariance model in the R^2 domain are plotted in **Figure 2B**, where the HFRS covariance model is analytically given by

$$c_{\widehat{X}}(\widehat{h}) = 0.75(1 - \frac{1.5\widehat{h}}{10^4} + \frac{0.5\widehat{h}^3}{10^{12}}) + 0.25 e^{-\frac{\widehat{h}}{3.5\times 10^5}}. \tag{12}$$

This covariance is also of reduced dimensionality compared to that of Equation (11). In addition, the theoretical covariance model of Equation (12) provided a better fit to the corresponding empirical covariance than the covariance model of Equation (11).

HFRS Incidence Mapping

The HFRS incidence maps were firstly generated using the P-BME technique. For illustration, the monthly HFRS incidence maps for the year 2006 (January–December) are shown in **Figure 3**. Additional HFRS incidence maps for all the years considered can be found in the section of "Supporting Information."

As regards spatial variation, three areas with considerable HFRS incidence are identified in the maps of **Figure 3**, particularly, in the eastern, the western and the southern parts of the Heilongjiang province. Among them, the eastern part shows high HFRS incidence values over a larger area than in the other two parts. As is noted in the section of Discussion, this happens because there exists a corresponding large area of croplands and rivers in the eastern part that are linked to increasing HFRS incidence. As regards the temporal HFRS variation in Heilongjiang province, the HFRS incidence variation exhibited two outbreaks within a year's time. Specifically, as is shown in **Figure 3** and Figure S12 (Supporting Information), the HFRS incidence begins to increase in April of 2006, then a peak is reached in June of 2006, and the HFRS incidence reduces significantly in September of 2006. The next outbreak is observed during September 2006 to February 2007 with the peak occurring in November of 2006. Interestingly, the number of HFRS cases during the autumn-winter period is much larger than those during the spring-summer period. As noted in the section of Discussion, the interpretation of this phenomenon is that it is probably due to the fact that the autumn-winter period coincides with the rice harvest season, i.e., after harvest the soil condition switches from a flood state to a dry state, which leads to the rodents dispersal or migration causing a higher number of infected cases).

In the western part of the Heilongjiang province, low HFRS incidence values were observed during the months February–September of the period 2006–2013 (but not for the year 2005). The HFRS incidence at the southern part of the Heilongjiang province remained high during the months of June, October, November, and December of each year considered. Apparently, HFRS is transmitted to the southern part of the province from its eastern part. Overall, a declining trend of HFRS incidence is observed in the maps of Figures S3–S11 (Supporting Information) for the period 2005–2013, which may be due, at least in part to the development of medical condition and disease prevention.

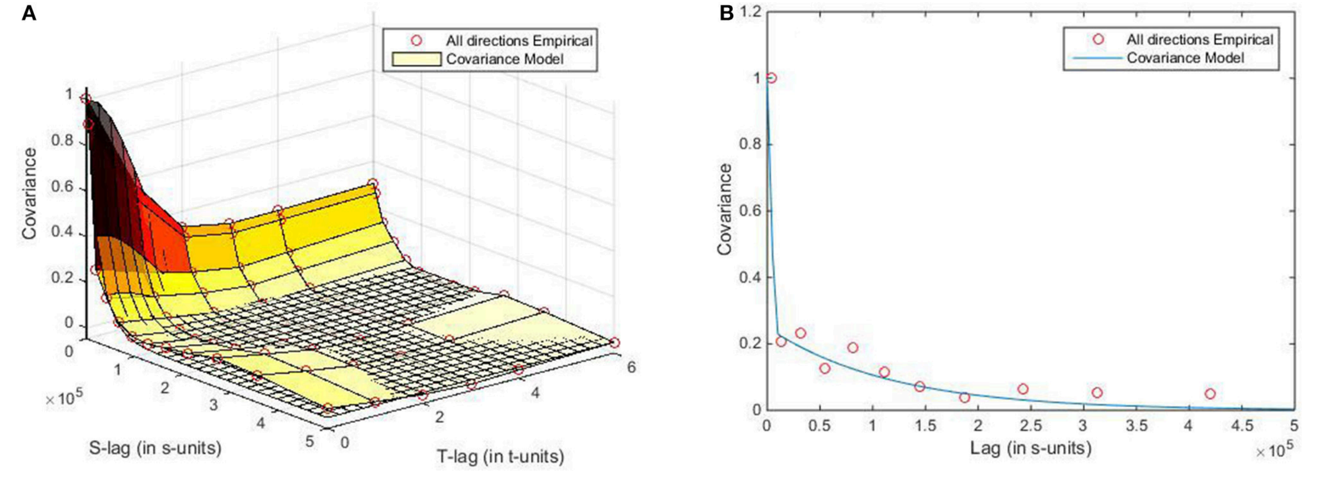


FIGURE 2 | Plots of the spatiotemporal empirical covariances and the fitted theoretical models in **(A)** the $R^2 \times T$ domain, and **(B)** the R^2 domain.

A cross-validation analysis of P-BME mapping technique vs. direct BME mapping is plotted in Figure S2 during the years 2005–2013: specifically, the P-BME mapping was more accurate in predicting HFRS at low incidence points, whereas the direct BME mapping was more accurate in predicting HFRS incidence at high incidence points during 2005 and 2007 (see, also, the months 1–36 of the time series of HFRS incidence in **Figure 4A**). Overall, the P-BME was on average a better predictor of the space-time HFRS incidence distribution in the Heilongjiang province than the direct BME: the mean absolute prediction error for BME over the entire domain was 0.524 cases/100,000 individuals, whereas that of P-BME was 0.459 cases/100,000 individuals. An explanation for the above results is given in the section of Discussion below.

Ensemble Empirical Mode Decomposition

Using the ensemble EMD method, five IMFs and one residual component were technically extracted from the HFRS and MEI series. The results are shown in **Figure 4**. In the first 40 months, the HFRS incidence series experienced much higher peaks than during the remaining months (**Figure 4A**), which is consistent with the findings of Figures S3–S11 (Supporting Information) above. The IMF frequency decreases from IMF1 to IMF5, whereas the corresponding IMF period increases. As it can be also seen, the long trends (residual component) of HFRS and MEI are decreasing.

Each IMF expresses a different fluctuation period (**Table 1**). For HFRS, the IMF1 to IMF5 represent incidence periods lasting 5.959, 8.936, 18.024, 39.002, and 64.004 months, respectively. The main HFRS inherent periods are 5.959, 8.936, and 18.024 months, according to the contribution percentage of IMF's variance. The corresponding MEI periods are 5.138, 17.662, 34.629, 50.129, and 108.466 months, respectively, whereas the main periods are 17.662, 34.629, and 50.129 months. By comparing the periods of each IMF, it was found that the IMF1 of HFRS has the same (6 months) period as the IMF1 of MEI. Similarly, the IMF3 of HFRS has the same (18 months) period with the IMF2 of MEI.

Hilbert–Huang and Marginal HFRS Incidence Spectra

The reason why the time-frequency technique (i.e., Hilbert–Huang transformation) is used at this study stage is that we seek to detect the similarities of the two series, in which case it can be concluded that the two series are inter-related. The Hilbert-Huang spectra of the HFRS and MEI series displayed in **Figures 5A,B** represent the time-frequency-energy distributions of the original series. As can be seen in **Figures 5A,B**, continuous instantaneous frequencies are detected in the low frequency region of the spectra (<0.15 cycle/month). The energy is also high in the low frequency region, especially for MEI, although some amounts of energy are detected in the middle and high frequency region of the spectra (>0.15 cycle/month).

Moreover, the energy associated with HFRS incidence is more discretely distributed than that associated with MEI. As is shown in **Figure 6**, the marginal spectra of HFRS and MEI have a similar peak occurring at ∼0.015–0.025 cycle/month. Also, we notice the peaks at 0.0694 and 0.0787 cycle/month. From 0 to 0.15 cycle/month, both spectra exhibit a decreasing trend. However, the MEI spectrum decreases rapidly to a small level and the energy is concentrated in this frequency range. Compared to the marginal MEI spectrum, the marginal HFRS spectrum presents a more complex fluctuation pattern across the entire frequency domain, and the amplitude (or energy) remains above the 0.15 cycle/month threshold.

Wavelet Coherency Analysis

As a result of the wavelet coherency analysis we obtained the coherency wavelet spectrum between the HFRS and MEI series shown in **Figure 7A**. This figure indicates that there exists a strong coherency between the two series with a periodicity of 6-months band, particularly, during the sampled months 22–31, 41–52, 75–80, and 85–94. A weaker coherency was also detected around the 12-month band during the sampled months 18∼28, 42∼52, and 61∼83.

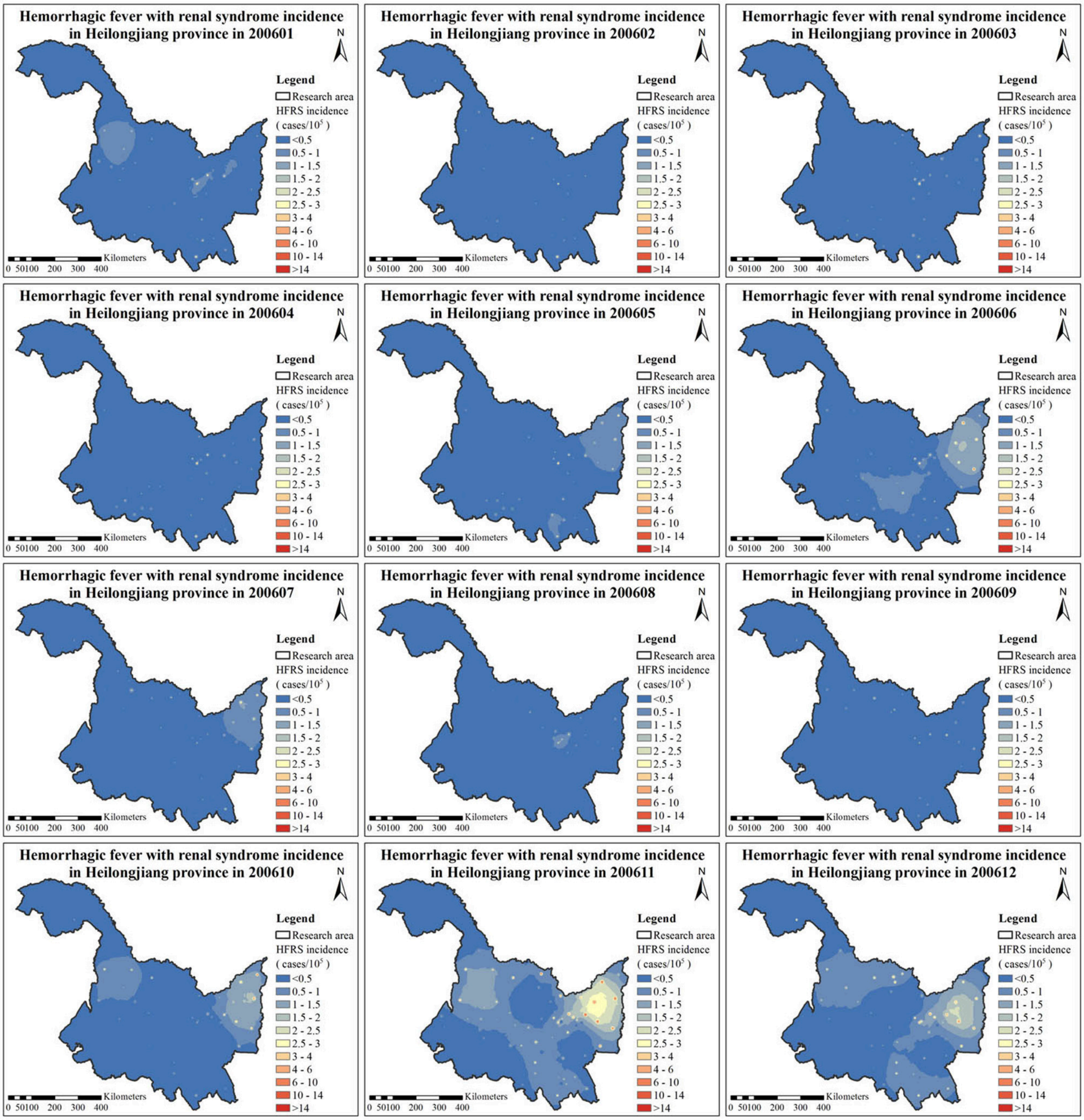


FIGURE 3 | HFRS incidence maps during the period Jan–Dec 2006.

The HFRS and MEI series reconstructed by means of the wavelet transformation are plotted in **Figures 7B,C**. These plots may be interpreted as providing an interesting demonstration of the HFRS and MEI series oscillations, which is useful for comparison purposes. A strong coherency is easily detected with 2 period month bands (i.e., 5–7 and 8–16 month band) during the sampled months mentioned above. This explains why the same oscillation pattern is found in **Figures 7B,C**.

DISCUSSION

Public health scientists and epidemiologists are increasingly in need of gaining insight about the space-time HFRS incidence variation, and about how climate change affects HFRS incidence dynamics. This is particularly true in Heilongjiang province, which is one of the most HFRS affected areas in China. To the best of our knowledge, very few studies have used

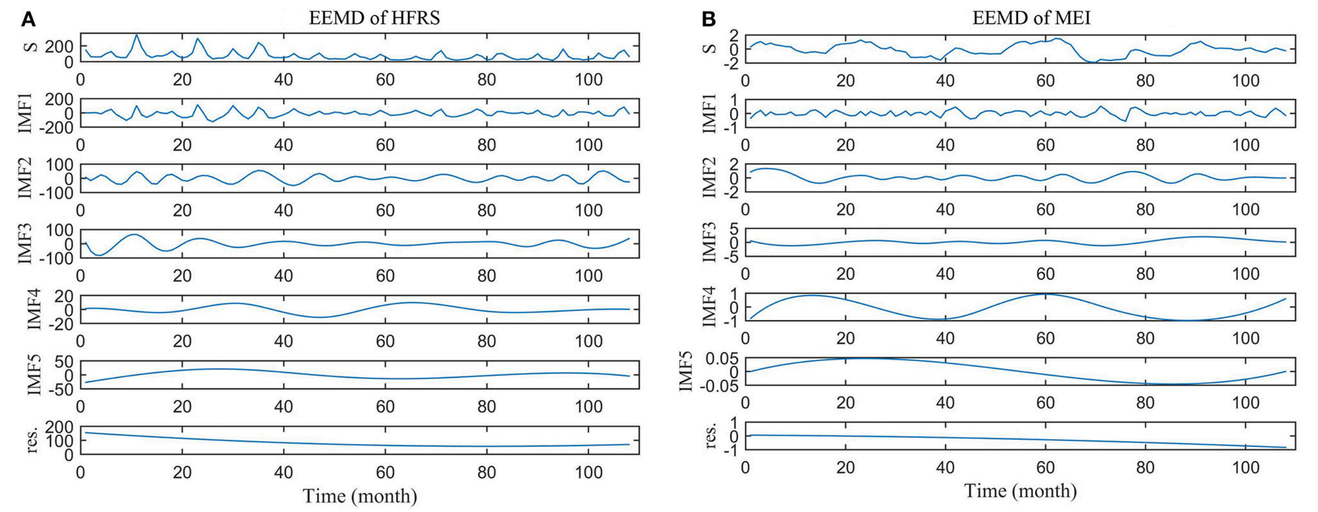


FIGURE 4 | Ensemble EMD of **(A)** HFRS series and **(B)** MEI series. S is the original series and res. denotes the residual series.

TABLE 1 | Statistics of ensemble EMD results.

Modes	IMF1	IMF2	IMF3	IMF4	IMF5	Res
HRFS						
Period (month)	5.959	8.936	18.024	39.002	64.004	–
Variance (%)	56.612	18.194	20.051	0.913	4.231	–
Correlation coefficient	0.607**	0.390**	0.271**	0.029	0.232*	0.381**
MEI						
Period (month)	5.138	17.662	34.629	50.129	108.466	–
Variance (%)	2.595	16.286	52.239	28.803	0.078	–
Correlation coefficient	0.172	0.436**	0.384**	0.340**	0.138	0.128

**Correlation is significant at the 0.05 level (2-tailed).*

***Correlation is significant at the 0.01 level (2-tailed).*

analytical methods to describe the space-time HFRS spread in Heilongjiang province. On the other hand, the associations between HFRS incidence and climate factors have been always assessed in terms of numerical modeling, for example, autoregressive integrated moving average models (ARIMA), seasonal ARIMA (SARIMA), ecological niche models (ENM), Poisson regression models, multiple regression, conditional logistic regression, and principal components regression (PCR) models [25, 34, 45–48]. Interestingly, none of these studies explored the association between HFRS incidence and climatic factors in the context of their co-variation.

Responding to the above need, the present work is a collaborative effort between the Zhejiang University (Zhoushan, China), the Institute of Disease Control and Prevention (Beijing, China), and the San Diego State University (California, USA). This collaboration led to the introduction of a combination of modern space-time modeling and mapping techniques from BME theory, Hilbert–Huang spectrum analysis and wavelet analysis in the study of the spatiotemporal HFRS incidence distribution in Heilongjiang province, and its association with climate change.

In particular, one of the main elements of this study is the implementation of the P-BME method [41] to analyze the space-time HFRS incidence spread in Heilongjiang province. Monthly HFRS incidence data were analyzed and processed across the Heilongjiang province during the period 2005–2013. Monthly HFRS data are used here because they lead to more accurate predictions than annual data (see below) and, also, they can serve better our goals to detect the temporal HFRS incidence pattern (which can be explained in ecological terms) and assess the association between the HFRS incidence pattern and climate change (MEI).

A key feature of the P-BME method is that technically it transfers the study of HFRS incidence spread from the original 3-D (i.e., two space dimensions plus time, $R^2 \times T$) domain onto a reduced dimensionality 2-D (i.e., two space dimensions, R^2) domain. In this way, the difficult to determine space-time distance (metric) is reduced to a much easier to define spatial distance, which means that the empirical space-time covariance of the monthly HFRS incidence distribution is accordingly transformed into a spatial covariance (see **Figure 2**). As a result, it is technically much easier to fit a theoretical model to the spatial than to the spatiotemporal empirical covariance of monthly HFRS data.

Next, for comparison purposes, the empirical covariances and the fitted theoretical models (in the $R^2 \times T$ and the R^2 domains) for the annual HFRS data are shown in Figure S13. This comparison shows that the temporal dependence of the annual HFRS data is much stronger than that of the monthly HFRS data (e.g., by comparing **Figure 2** and Figure S13, it is seen that the annual HFRS covariance value at time lag $\tau = 4$ is about 0.7 $(cases/10^5)^2$ compared to 0.3 $(cases/10^5)^2$ for the monthly covariance).

As regards the space-time mapping of HFRS spread in Heilongjiang province, it was found that the weaker the temporal dependence of the HFRS empirical covariance is, the better is the

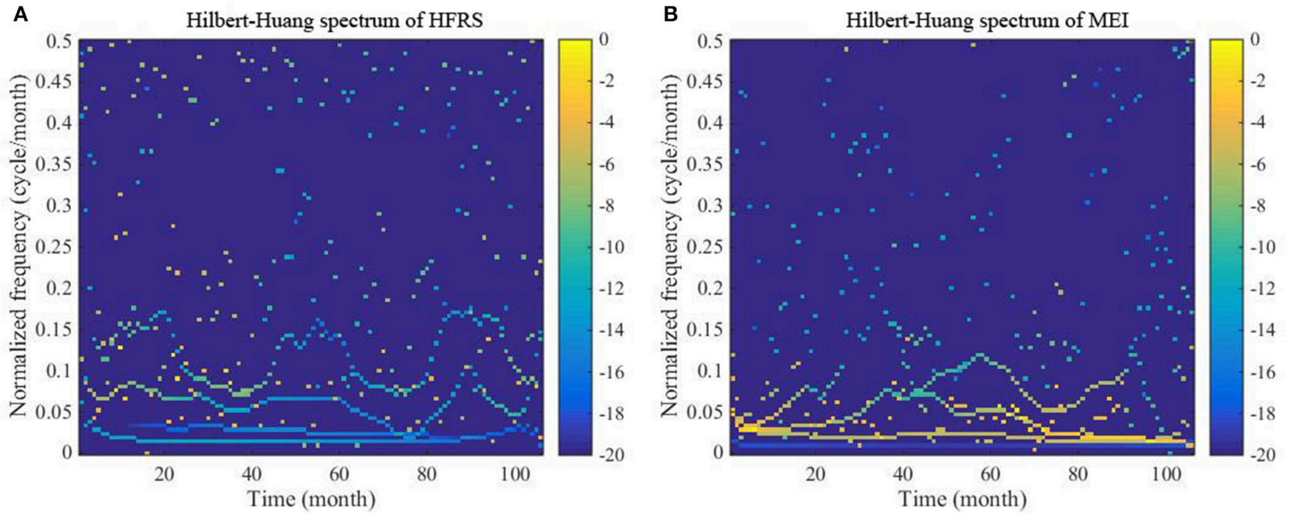


FIGURE 5 | Hilbert-Huang spectra of **(A)** the HFRS series and **(B)** the MEI series. The color bar ranging from dark blue to yellow indicates energy variation from minimum to maximum.

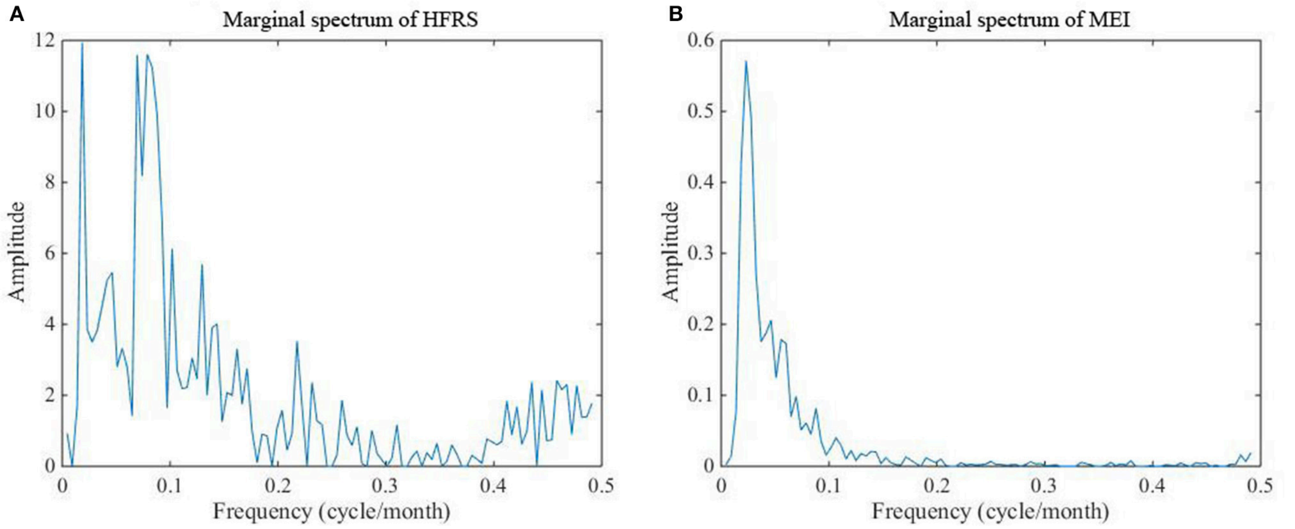


FIGURE 6 | Marginal spectra of **(A)** the HFRS series and **(B)** the MEI series.

P-BME performance compared to that of the direct BME method. This improved performance of P-BME in this case is explained by the fact that since the time argument is technically imbedded within the projected coordinates of the reduced dimensionality domain, the temporal points with the stronger dependence (compared to the spatial ones) are not explicitly taken into account, instead, only traveling spatial points are considered in HFRS prediction and mapping, and the HFRS prediction error would be larger in this case. Compared to the original BME method, the P-BME was found to provide on average more accurate HFRS incidence predictions in the case of monthly HFRS incidence data (the mean absolute prediction error for P-BME is 0.459 cases/100,000 individuals vs. 0.524 cases/100,000 individuals for direct BME), whereas the opposite was the case for annual HFRS incidence data (the mean absolute prediction error for direct BME was 3.34 cases/100,000 individuals and for P-BME it was 4.56 cases/100,000 individuals).

Other findings of this study included the following. Three core areas were observed in the HFRS incidence distribution maps we obtained for the period Jan 2005–Dec 2013, particularly, the eastern, western and southern parts of Heilongjiang province. As the drainage map of Heilongjiang province shows (Figure S14), the Wusuli and Songhua rivers, as well as parts of the Heilong river belong to the eastern part, and the Neng and Mudan rivers belong to western and southern parts, respectively, of the Heilongjiang Province. Regarded as major water sources, these

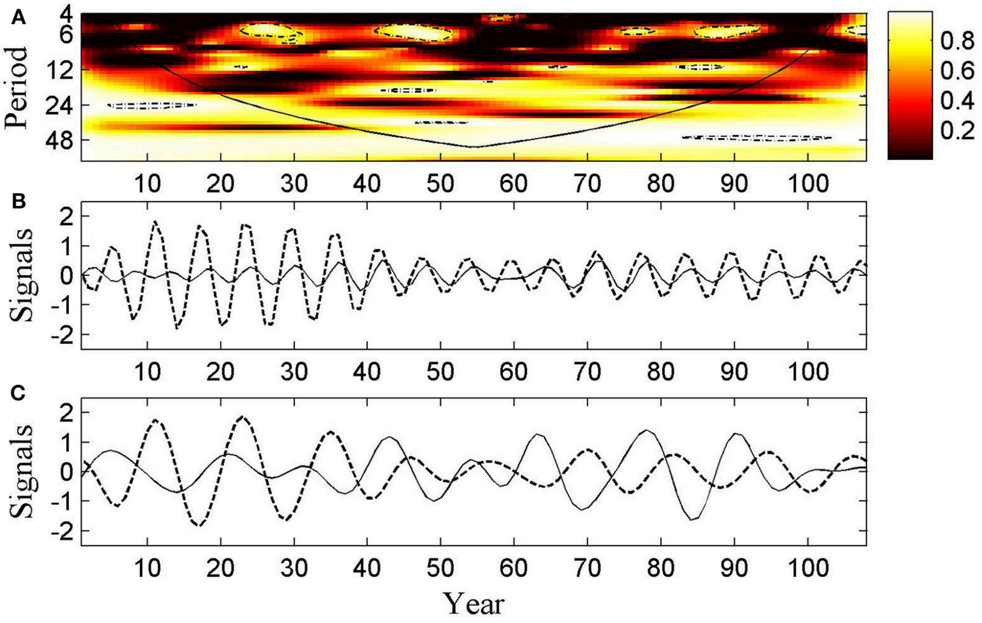


FIGURE 7 | The HFRS-MEI association. **(A)** Wavelet coherency between HFRS and MEI series. The colors depict coherency values from black (0) to white (1); black line represents the cone of influence that delimits the region that is not influenced by edge effects; the dash dot line shows $a = 5\%$ significance level computed based on 500 simple bootstrap. **(B)** The oscillation series during the period 5–7 months band that was reconstructed from the real part of the wavelet transformed series. **(C)** The oscillation series during the period 8–16 months band reconstructed from the real part of the wavelet transformed series. Dash lines represent oscillation of the HFRS series, while black line represents oscillation of the MEI series.

river basins provide a suitable environment for rodent hosts and their reproduction. The presence of a river or a pond can be a risk for human HFRS infection [49]. Bao et al. [50] found that HFRS incidence had a strong correlation with distance to rivers (in particular, HFRS incidence had a quadratic relationship with distance to rivers, and $R^2 = 0.999$, $p = 0.000$). On the other hand, severe droughts can significantly decrease HFRS incidence [2, 51]. As rivers can provide sufficient water for vegetation irrigation purposes, the croplands are always located near rivers, and Heilongjiang province is no exception.

In the Heilongjiang province, croplands, mixed forests, and cropland/natural vegetation mosaic account for 38.98, 26.29, and 16.92% of the territory, respectively (Figure S15). Specifically, croplands are largely distributed in the eastern and western parts of Heilongjiang province, and there are also some croplands in the southern part of Heilongjiang province, which correspond to the three core HFRS areas mentioned above. It has been found that crop production is highly correlated with HFRS incidence with a correlation coefficient $r = 0.96$ ($p = 0.005$) due to the fact that crop can directly or indirectly serve as food for rodent hosts [45]. Increasing food availability contributes to the growth of rodent host population. This can raise the infection probability of humans, especially farmers, who have a higher likelihood to come in contact with these animals [52, 53]. Farmers usually don't have steady jobs other than farming, and they also have higher mobility compared to other professions. Therefore, they may carry and spread hantaviruses to wider areas than the croplands during their traveling after the farming season. Under these conditions, HFRS may spread rapidly. Being aware of the above high HFRS spread likelihood, it is necessary to implement public health interventions in the core areas of the Heilongjiang province to avoid HFRS outbreaks and spread (these interventions include, e.g., the extermination of potential rodent hosts and vaccination).

In addition, June and November HFRS incidence peaks were found during the years 2005–2013 (Figure S12). Interestingly, this bimodal temporal pattern was also found in Hubei province after 1995, as result of the Seoul virus-related HFRS spring outbreaks and the Hantaan virus-related HFRS winter outbreaks [54]. Another study found that HFRS associated with wild and house rodents occur during different seasons [55]. In view of the fact that land-use can affect virus occurrence in hosts by influencing movement and contact rate [56, 57], we notice that Heilongjiang is located in a high latitude area, where paddy rice only grows once a year (during May and October). At the beginning of sowing season in May, soils are irrigated to remain in a flood state for paddy rice growth, and the soil conditions change from drying to flooding. After harvesting, the soil will return to drying conditions in October. These switches between different land environmental conditions may cause a cascade of factors contributing to infectious disease emergence, especially invasive alien species due to the fact that they are invaders over their natural range in the new environment [58–60]. More specifically, an agricultural ecosystem is particularly vulnerable to invasive alien species and anthropogenic activities that can initiate or accelerate the introduction or invasion of alien species

[58]. Sufficient food availability during June contributes to rodent reproduction, whereas insufficient food availability or the dry conditions of November will result in spur sudden dispersal or migration events, both of which can increase HFRS infection [61]. Whats more, the edges of paddy fields may involve ecotones as habitats with infectious disease and animal reservoir hosts being abundant in wildlife [62]. Following a month-long period of environmental change, HFRS outbreaks may occur.

Generally, the study of real world phenomena is highly complex and interdisciplinary, in particular the study of large spatial scale climate and disease variation (which can be affected by biological, social, geographic, economic, medical factors etc.; [63]). Hence, the observation series may contain a large amount of direct and indirect information. With such complex information, it can be really hard for scientists to collect data from various disciplines and explore the relationship among them through hypothesis- and equation-driven methods.

In view of the above considerations, time-frequency analysis methods, regarded as data-mining tools, constitutes another major component of the present study. The results can reveal the intrinsic variation patterns of HFRS incidence and MEI series, as well as the dynamic characteristics of the HFRS and MEI cycles in the time-frequency domain. Understanding the association between HFRS incidence and climate change (using MEI as a proxy, measuring coupled oceanic-atmospheric character of ENSO event) provides a potential auxiliary way to assess the public health effects of global climate change, since climate variability has important effects on wildlife population dynamics [64, 65]. The Hilbert-Huang transformation is a powerful tool for solving mode-mixing problems and can be also used as a filter for decomposing raw HFRS incidence series into several independent series with disparate modes, i.e., IMFs [66]. Different component series were obtained that describe various inherent disease characteristics that cannot be detected in the raw series. Our results showed that both HFRS and MEI series have six types of characteristic components. A monotonic declining trend is shown in **Figure 4**, and both series are characterized by 6- and 18-months periods, approximately (**Table 1**), indicating that similar patterns are hidden in the variation features of the two series. For further analysis, the Hilbert–Huang and the marginal spectra were used to assess the strength of series variation in the combined time-frequency domain. For both series, stable and consistent variations were observed at low frequency regions, although certain discrete fluctuations can be found in the HFRS spectra that are not observed in the MEI spectra. Such differences may not be explained by climate change (MEI) but rather in terms of non-climatic factors, e.g., population immunity, public health condition, and socio-economic factors [10, 54, 67].

Moreover, wavelet coherency analysis showed that the HFRS incidence has a strong association with MEI during a 6-month period (**Figure 7**). These results suggest that the HFRS incidence dynamics are interrelated with climate change and the MEI can serve as a potential predictor of HFRS occurrence. We notice that similar results have been found for diseases like dengue fever, dengue hemorrhagic fever, hantavirus cardiopulmonary syndrome, and malaria [68–71]. Moreover, regional precipitation is known to be influenced by ENSO, showing the strongest interrelation with climate variability around the Globe [72]. Increasing precipitation provides sufficient soil moisture for improving ecosystem productivity [64]. As a result, the number of rodents grows rapidly, leading to increasing contact rates between rodents and between rodents and humans [61]. The HFRS infection rate increases under the above ecological changes. A deeper understanding of the association between climate change and HFRS incidence can provide a potential tool of early HFRS outbreak warning, especially concerning short-term effects.

Certain limitations of the present work should be acknowledged, which are rather typical for this kind of quantitative studies. The first one is data limitation, i.e., the HFRS dataset used is an aggregated set that does not distinguish the infectious HFRS types, e.g., Hantaan virus or Seoul virus, for which the infectious dynamics may be different. Second, the impact of climate change on human HFRS incidence can be twofold: impact from rodents to rodents and infection from rodents to humans. Therefore, distinct studies of these two possibilities would offer a better understanding of HFRS transmission. Third, some other impact factors could also be included in HFRS pattern analysis, such as population immunity and socio-economic factors. Future work should focus on a more detailed analysis of spatiotemporal intensity differences of the various environmental factors impacting HFRS incidence.

In sum, the present work provides a quantitative study of the HFRS incidence spread in Heilongjiang province (China). Three core areas with high HFRS incidences were identified. In addition, time-frequency analysis provides evidence that HFRS incidence is closely related to climate change.

AUTHOR CONTRIBUTIONS

All authors listed have made a substantial, direct and intellectual contribution to the work, and approved it for publication.

ACKNOWLEDGMENTS

The authors acknowledge with appreciation the Matlab codes (processing wavelet analysis) provided by Professor Bernard Cazelles. The work was supported by grants from the National Natural Science Foundation of China (No. 41671399 and No. 11501339).

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: http://journal.frontiersin.org/article/10.3389/fams.2017.00016/full#supplementary-material

REFERENCES

1. Yan L, Fang LQ, Huang HG, Zhang LQ, Feng D, Zhao WJ, et al. Landscape elements and Hantaan virus–related hemorrhagic fever with renal syndrome, People's Republic of China. *Emerg Infect Dis.* (2007) **13**:1301–6. doi: 10.3201/eid1309.061481
2. Tian H, Yu P, Bjørnstad ON, Cazelles B, Yang J, Tan H, et al. Anthropogenically driven environmental changes shift the ecological dynamics of hemorrhagic fever with renal syndrome. *PLoS Pathog.* (2017) **13**:e1006198. doi: 10.1371/journal.ppat.1006198
3. Liu J, Xue FZ, Wang JZ, Liu QY. Association of haemorrhagic fever with renal syndrome and weather factors in Junan County, China: a case-crossover study. *Epidemiol Infect.* (2013) **141**:697–705. doi: 10.1017/S0950268812001434
4. Chinikar S, Javadi A, Hajiannia A, Ataei B, Jalali T, Khakifirouz S, et al. First evidence of Hantavirus in central Iran as an emerging viral disease. *Adv Infect Dis.* (2014) **4**:173–7. doi: 10.4236/aid.2014.44024
5. Simmons JH, Riley LK. Hantaviruses: an overview. *Compar Med.* (2002) **52**:97–110. Available online at: http://www.ingentaconnect.com/content/aalas/cm/2002/00000052/00000002/art00003
6. Krautkrämer E, Zeier M, Plyusnin A. Hantavirus infection: an emerging infectious disease causing acute renal failure. *Kidney Int.* (2012) **83**:23–7. doi: 10.1038/ki.2012.360
7. Jiang H, Du H, Wang L, Wang P, Bai X. Hemorrhagic fever with renal syndrome: pathogenesis and clinical picture. *Front Cell Infect Microbiol.* (2016) **6**:1. doi: 10.3389/fcimb.2016.00001
8. Fang L, Yan L, Liang S, de Vlas SJ, Feng D, Han X, et al. Spatial analysis of hemorrhagic fever with renal syndrome in China. *BMC Infect Dis.* (2006) **6**:77. doi: 10.1186/1471-2334-6-77
9. Zou LX, Chen MJ, Sun L. Haemorrhagic fever with renal syndrome: literature review and distribution analysis in China. *Int J Infect Dis.* (2016) **43**:95–100. doi: 10.1016/j.ijid.2016.01.003
10. Zhang YZ, Zou Y, Fu ZF, Plyusnin A. Hantavirus Infections in Humans and Animals, China. *Emerg Infect Dis.* (2010) **16**:1195–203. doi: 10.3201/eid1608.090470
11. Fidan K, Polat M, Isıyel E, Kalkan G, Tezer H, Söylemezoğlu O. An adolescent boy with acute kidney injury and fever: answers. *Pediatr Nephrol.* (2013) **28**:2115–6. doi: 10.1007/s00467-012-2360-0
12. Dearing MD, Dizney L. Ecology of hantavirus in a changing world. *Ann NY Acad Sci.* (2010) **1195**:99–112. doi: 10.1111/j.1749-6632.2010.05452.x
13. Pépin M, Dupinay T, Zilber AL, McElhinney LM. Of rats and pathogens: pathogens transmitted by urban rats with an emphasis on hantaviruses. *CAB Rev.* (2016) **11**:1–13. doi: 10.1079/PAVSNNR201611019
14. Li S, Ren H, Hu W, Lu L, Xu X, Zhuang D, et al. Spatiotemporal heterogeneity analysis of hemorrhagic Fever with renal syndrome in china using geographically weighted regression models. *Int J Environ Res Public Health* (2014) **11**:12129–47. doi: 10.3390/ijerph111212129
15. Zhang S, Wang S, Yin W, Liang M, Li J, Zhang Q, et al. Epidemic characteristics of hemorrhagic fever with renal syndrome in China, 2006–2012. *BMC Infect Dis.* (2014) **14**:384. doi: 10.1186/1471-2334-14-384
16. Zhang WY, Wang LY, Liu YX, Yin WW, Hu WB, Magalhaes RJ, et al. Spatiotemporal transmission dynamics of hemorrhagic fever with renal syndrome in China, 2005–2012. *PLoS Negl Trop Dis.* (2014) **8**:e3344. doi: 10.1371/journal.pntd.0003344
17. Christakos G. A Bayesian/maximum-entropy view to the spatial estimation problem. *Math Geol.* (1990) **22**:763–76. doi: 10.1007/BF00890661
18. Christakos G. Spatiotemporal information systems in soil and environmental sciences. *Geoderma* (1998) **85**:141–79. doi: 10.1016/S0016-7061(98)00018-4
19. Christakos G. *Modern Spatiotemporal Geostatistics*. New York, NY: Oxford University Press (2000).
20. He, J, Kolovos, A. Bayesian maximum entropy approach and its applications: a review. *Stoch Environ Res Risk Assess.* (2017) doi: 10.1007/s00477-017-1419-7. [Epub ahead of print].
21. Law DCG, Bernstein KT, Serre ML, Schumacher CM, Leone PA, Rompalo AM, et al. Modeling a syphilis outbreak through space and time using the Bayesian maximum entropy approach. *Ann Epidemiol.* (2006) **16**:797–804. doi: 10.1016/j.annepidem.2006.05.003
22. Choi KM, Yu HL, Wilson ML. Spatiotemporal statistical analysis of influenza mortality risk in the State of California during the period 1997–2001. *Stoch Environ Res Risk Assess.* (2008) **22**:15–25. doi: 10.1007/s00477-007-0168-4
23. Clement J, Vercauteren J, Verstraeten WW, Ducoffre G, Barrios JM, Van Ranst A-M, et al. Relating increasing hantavirus incidences to the changing climate: the mast connection. *Int J Health Geogr.* (2009) **8**:1. doi: 10.1186/1476-072X-8-1
24. Zhang WY, Guo WD, Fang LQ, Li CP, Bi P, Glass GE, et al. Climate variability and hemorrhagic fever with renal syndrome transmission in Northeastern China. *Environ Health Perspect.* (2010) **118**:915–20. doi: 10.1289/ehp.0901504
25. Liu X, Jiang B, Gu W, Liu Q. Temporal trend and climate factors of hemorrhagic fever with renal syndrome epidemic in Shenyang City, China. *BMC Infect Dis.* (2011) **11**:331. doi: 10.1186/1471-2334-11-331
26. Xiao H, Tian HY, Cazelles B, Li XJ, Tong SL, Gao LD, et al. Atmospheric moisture variability and transmission of hemorrhagic fever with renal syndrome in Changsha City, Mainland China, 1991–2010. *PLoS Negl Trop Dis.* (2013) **7**:e2260. doi: 10.1371/journal.pntd.0002260
27. Tian HY, Yu PB, Luis AD, Bi P, Cazelles B, Laine M, et al. Changes in rodent abundance and weather conditions potentially drive hemorrhagic fever with renal syndrome outbreaks in Xi'an, China, 2005–2012. *PLoS Negl Trop Dis.* (2015) **9**:e0003530. doi: 10.1371/journal.pntd.0003530
28. Huang NE, Shen Z, Long SR, Wu MC, Shih HH, Zheng Q, et al. The empirical mode decomposition and the Hilbert spectrum for nonlinear and non-stationary time series analysis. *Proc R Soc A* (1998) **454**:903–95. doi: 10.1098/rspa.1998.0193
29. Huang NE, Wu ML, Qu W, Long SR, Shen SS. Applications of Hilbert–Huang transform to non-stationary financial time series analysis. *Appl Stoch Models Bus Indus.* (2003) **19**:245–68. doi: 10.1002/asmb.501
30. Huang NE, Wu Z. A review on Hilbert-Huang transform: method and its applications to geophysical studies. *Rev Geophys.* (2008) **46**:RG2006. doi: 10.1029/2007RG000228
31. Grinsted A, Moore JC, Jevrejeva S. Application of the cross wavelet transform and wavelet coherence to geophysical time series. *Nonlin Process Geophys.* (2004) **11**:561–6. doi: 10.5194/npg-11-561-2004
32. Thai KT, Cazelles B, Van Nguyen N, Vo LT, Boni MF, Farrar J, et al. Dengue dynamics in Binh Thuan province, southern Vietnam: periodicity, synchronicity and climate variability. *PLoS Negl Trop Dis.* (2010) **4**:e747. doi: 10.1371/journal.pntd.0000747
33. Chowell G, Cazelles B, Broutin H, Munayco CV. The influence of geographic and climate factors on the timing of dengue epidemics in Perú, 1994-2008. *BMC Infect Dis.* (2011) **11**:164. doi: 10.1186/1471-2334-11-164
34. Li CP, Cui Z, Li SL, Magalhaes RJS, Wang BL, Zhang C, et al. Association between hemorrhagic fever with renal syndrome epidemic and climate factors in Heilongjiang Province, China. *Am J Trop Med Hyg.* (2013) **89**:1006–12. doi: 10.4269/ajtmh.12-0473
35. Curran LM, Caniago I, Paoli GD, Astianti D, Kusneti, M, Haeruman H. Impact of El niño and logging on canopy tree recruitment in borneo. *Science* (1999) **286**:2184–8. doi: 10.1126/science.286.5447.2184
36. Wolter K, Timlin MS. El Niño/Southern oscillation behaviour since 1871 as diagnosed in an extended multivariate ENSO index (MEI. ext). *Int J Climatol.* (2011) **31**:1074–87. doi: 10.1002/joc.2336
37. Christakos G. *Spatiotemporal Random Fields: Theory and Applications*. Amsterdam: Elsevier (2017).
38. Reyes JM, Serre ML. An LUR/BME framework to estimate PM2. 5 explained by on road mobile and stationary sources. *Environ Sci Technol.* (2014) **48**:1736–44. doi: 10.1021/es4040528
39. Kyriakidis PC, Journel AG. Geostatistical space–time models: a review. *Math Geol.* (1999) **31**:651–84. doi: 10.1023/A:1007528426688
40. Stein ML. Space–time covariance functions. *J Am Stat Assoc.* (2005) **100**:310–21. doi: 10.1198/016214504000000854
41. Christakos G, Zhang C, He J. A traveling epidemic model of space–time disease spread. *Stoch Environ Res Risk Assess.* (2017) **31**:305–14. doi: 10.1007/s00477-016-1298-3
42. Wu Z, Huang N. Ensemble empirical mode decomposition: a noise-assisted data analysis method. *Adv Adapt Date Anal.* (2009) **01**:1–41. doi: 10.1142/S1793536909000047
43. Ju K, Zhou D, Zhou P, Wu J. Macroeconomic effects of oil price shocks in China: an empirical study based on Hilbert–Huang transform and

event study. *Appl Energy* (2014) **136**:1053–66. doi: 10.1016/j.apenergy.2014.08.037

44. Cazelles B, Chavez M, Berteaux D, Ménard F, Vik JO, Stenseth NC, et al. Wavelet analysis of ecological time series. *Oecologia* (2008) **156**:287–304. doi: 10.1007/s00442-008-0993-2
45. Bi P, Wu X, Zhang F, Parton KA, Tong S. Seasonal rainfall variability, the incidence of hemorrhagic fever with renal syndrome, and prediction of the disease in low-lying areas of China. *Am J Epidemiol.* (1998) **148**:276–81. doi: 10.1093/oxfordjournals.aje.a009636
46. Xiao H, Gao L, Li X, Lin X, Dai X, Zhu P, et al. Environmental variability and the transmission of haemorrhagic fever with renal syndrome in Changsha, People's Republic of China. *Epidemiol Infect.* (2013) **141**:1867–75. doi: 10.1017/S0950268812002555
47. Xiao H, Lin X, Gao L, Huang C, Tian H, Li N, et al. Ecology and geography of hemorrhagic fever with renal syndrome in Changsha, China. *BMC Infect Dis.* (2013) **13**:305. doi: 10.1186/1471-2334-13-305
48. Xiao H, Liu HN, Gao LD, Huang CR, Li Z, Lin XL, et al. Investigating the effects of food available and climatic variables on the animal host density of hemorrhagic fever with renal syndrome in Changsha, China. *PLoS ONE* (2013) **8**:e61536. doi: 10.1371/journal.pone.0061536
49. Li J, Chen Z, Hou T, Xing Y, Cai Z, Yang Y. Study on the risk factors of hemorrhagic fever with renal syndrome in Xi'an City (in Chinese). *Chin J Dis Control Prevent.* (2013) **17**:564–6. Available online at: http://www.cqvip.com/qk/80677a/201307/46406708.html
50. Bao C, Liu W, Zhu Y, Liu W, Hu J, Liang Q, et al. The spatial analysis on hemorrhagic fever with renal syndrome in Jiangsu Province, China based on geographic information system. *PLoS ONE* (2014) **9**:e83848. doi: 10.1371/journal.pone.0083848
51. Yang H, Xia X, Wu Y. Analysis of trend and impact factors of hemorrhagic fever with renal syndrome in Xi'an City, 1999-2000 (in Chinese). In: *First Annual Academic Conference of Chinese Preventive Medicine Association* (Jinan) (2002), p. 330–1. Available online at: http://cpfd.cnki.com.cn/Article/CPFDTOTAL-ZHYF200208001678.htm
52. Ruo SL, Li YL, Tong Z, Ma QR, Liu ZL, Fisher-Hoch SP. Retrospective and prospective studies of hemorrhagic fever with renal syndrome in rural China. *J Infect Dis.* (1994) **170**:527–34. doi: 10.1093/infdis/170.3.527
53. Engelthaler DM, Mosley DG, Cheek JE, Levy CE, Komatsu KK, Frampton JW. Climatic and environmental patterns associated with hantavirus pulmonary syndrome, Four Corners region, United States. *Emerg Infect Dis.* (1999) **5**:87. doi: 10.3201/eid0501.990110
54. Zhang YH, Ge L, Liu L, Huo XX, Xiong HR, Liu YY, et al. The epidemic characteristics and changing trend of hemorrhagic fever with renal syndrome in Hubei Province, China. *PLoS ONE* (2014) **9**:e92700. doi: 10.1371/journal.pone.0092700
55. Chen HX, Qiu FX, Dong BJ, Ji SZ, Li YT, Wang Y, et al. Epidemiological studies on hemorrhagic fever with renal syndrome in China. *J Infect Dis.* (1986) **154**:394–8. doi: 10.1093/infdis/154.3.394
56. Klempa B. Hantaviruses and climate change. *Clin Microbiol Infect.* (2009) **15**:518–23. doi: 10.1111/j.1469-0691.2009.02848.x
57. Zhang W-Y, Fang L-Q, Jiang J-F, Hui F-M, Glass GE, Yan L, et al. Predicting the risk of hantavirus infection in Beijing, People's Republic of China. *Am J Trop Med Hyg.* (2009) **80**:678–83. doi: 10.4269/ajtmh.2009.80.678
58. McNeely JA, Mooney HA, Neville LE, Schei PJ, Waage J (eds.). *Global Strategy on Invasive Alien Species.* Gland: IUCN (2001).
59. Patz JA, Daszak P, Tabor GM, Aguirre AA, Pearl M, Epstein J, et al. Unhealthy landscapes: policy recommendations on land use change and infectious disease emergence. *Environ Health Perspect.* (2004) **112**:1092–8. doi: 10.1289/ehp.6877
60. Tian H, Huang S, Zhou S, Bi P, Yang Z, Li X, et al. Surface water areas significantly impacted 2014 dengue outbreaks in Guangzhou, China. *Environ Res.* (2016) **150**:299–305. doi: 10.1016/j.envres.2016.05.039
61. Parmenter RR, Yadav EP, Parmenter CA, Ettestad P, Gage KL. Incidence of plague associated with increased winter-spring precipitation in New Mexico. *Am J Trop Med Hyg.* (1999) **61**:814–21. doi: 10.4269/ajtmh.1999.61.814
62. Despommier D, Ellis BR, Wilcox BA. The role of ecotones in emerging infectious diseases. *Ecohealth* (2006) **3**:281–9. doi: 10.1007/s10393-006-0063-3
63. Moustakas A. Spatio-temporal data mining in ecological and veterinary epidemiology. *Stoch Environ Res Risk Assess.* (2017) **31**:829–34. doi: 10.1007/s00477-016-1374-8
64. Stenseth NC, Mysterud A, Ottersen G, Hurrell JW, Chan KS, Lima M. Ecological effects of climate fluctuations. *Science* (2002) **297**:1292–6. doi: 10.1126/science.1071281
65. Patz JA, Campbell-Lendrum D, Holloway T, Foley JA. Impact of regional climate change on human health. *Nature* (2005) **438**:310–7. doi: 10.1038/nature04188
66. Tsai FF, Fan SZ, Lin YS, Huang NE, Yeh JR. Investigating power density and the degree of nonlinearity in intrinsic components of anesthesia EEG by the hilbert-huang transform: an example using ketamine and alfentanil. *PLoS ONE* (2016) **11**:e0168108. doi: 10.1371/journal.pone.0168108
67. Song G. Epidemiological progresses of hemorrhagic fever with renal syndrome in China. *Chin Med J.* (1999) **112**:472–7.
68. Hay SI, Myers MF, Burke DS, Vaughn DW, Endy T, Rogers DJ. Etiology of interepidemic periods of mosquito-borne disease. *Proc Natl Acad Sci USA.* (2000) **97**:9335–9. doi: 10.1073/pnas.97.16.9335
69. Hjelle B, Glass GE. Outbreak of hantavirus infection in the four corners region of the United States in the wake of the 1997–1998 El Nino—Southern oscillation. *J Infect Dis.* (2000) **181**:1569–73. doi: 10.1086/315467
70. Zhou G, Minakawa N, Githeko AK, Yan G. Association between climate variability and malaria epidemics in the East African highlands. *Proc Natl Acad Sci USA.* (2004) **101**:2375–80. doi: 10.1073/pnas.0308714100
71. Tipayamongkholgul M, Fang CT, Klinchan S, Liu CM, King CC. Effects of the El Niño-Southern Oscillation on dengue epidemics in Thailand, 1996-2005. *BMC Public Health* (2009) **9**:422. doi: 10.1186/1471-2458-9-422
72. Ropelewski CF, Halpert MS. Global and regional scale precipitation patterns associated with the El Niño/Southern Oscillation. *Monthly Weather Rev.* (1987) **115**:1606–26.

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers
in Ecology and Evolution

HYPOTHESIS AND THEORY
published: 07 August 2017
doi: 10.3389/fevo.2017.00091

Evaluating Hypotheses of Plant Species Invasions on Mediterranean Islands: Inverse Patterns between Alien and Endemic Species

Alexander Bjarnason[1]*, *Stelios Katsanevakis*[2], *Alexandros Galanidis*[3], *Ioannis N. Vogiatzakis*[4] *and Aristides Moustakas*[1]*

[1] School of Biological and Chemical Sciences, Queen Mary University of London, London, United Kingdom, [2] Department of Marine Sciences, University of the Aegean, Mytilene, Greece, [3] Department of Environment, University of the Aegean, Mytilene, Greece, [4] School of Pure and Applied Sciences, Open University of Cyprus, Nicosia, Cyprus

Invasive alien species cause major changes to ecosystem functioning and patterns of biodiversity, and the main factors involved in invasion success remain contested. Using the Mediterranean island of Crete, Greece as a case study, we suggest a framework for analyzing spatial data of alien species distributions, based on environmental predictors, aiming to gain an understanding of their spatial patterns and spread. Mediterranean islands are under strong ecological pressure from invading species due to their restricted size and increased human impact. Four hypotheses of invasibility, the "propagule pressure hypothesis" (H1), "biotic resistance hypothesis vs. acceptance hypothesis" (H2), "disturbance-mediated hypothesis" (H3), and "environmental heterogeneity hypothesis" (H4) were tested. Using data from alien, native, and endemic vascular plant species, the propagule pressure, biotic resistance vs. acceptance, disturbance-mediated, and environmental heterogeneity hypotheses were tested with Generalized Additive Modeling (GAM) of 39 models. Based on model selection, the optimal model includes the positive covariates of native species richness, the negative covariates of endemic species richness, and land area. Variance partitioning between the four hypotheses indicated that the biotic resistance vs. acceptance hypothesis explained the vast majority of the total variance. These results show that areas of high species richness have greater invasibility and support the acceptance hypothesis and "rich-get-richer" distribution of alien species. The negative correlation between alien and endemic species appears to be predominantly driven by altitude, with fewer alien and more endemic species at greater altitudes, and habitat richness. The negative relationship between alien and endemic species richness provides potential for understanding patterns of endemic and alien species on islands, contributing to more effective conservation strategies.

Keywords: biological invasions, alien species, endemic species, statistical hypothesis testing, data mining, statistical interpretation, island ecosystems, Crete

Edited by:
Juergen Pilz,
Alpen-Adria-Universität Klagenfurt,
Austria

Reviewed by:
Ijaz Hussain,
Quaid-i-Azam University, Pakistan
Luis Gomez,
University of Las Palmas de Gran
Canaria, Spain

***Correspondence:**
Alexander Bjarnason
bt03175@qmul.ac.uk
Aristides Moustakas
arismoustakas@gmail.com

Specialty section:
This article was submitted to
Environmental Informatics,
a section of the journal
Frontiers in Ecology and Evolution

Received: *25 April 2017*
Accepted: *21 July 2017*
Published: *07 August 2017*

Citation:
Bjarnason A, Katsanevakis S, Galanidis A, Vogiatzakis IN and Moustakas A (2017) Evaluating Hypotheses of Plant Species Invasions on Mediterranean Islands: Inverse Patterns between Alien and Endemic Species. Front. Ecol. Evol. 5:91. doi: 10.3389/fevo.2017.00091

INTRODUCTION

Alien species are non-indigenous taxa introduced to areas beyond their natural distribution and biogeographical barriers by human activity. They can become invasive and have major impacts on native biodiversity by causing local or global extinctions, modifying community composition, and impeding or altering ecosystem functioning (Kolar and Lodge, 2001; Richardson and Pyšek, 2006; Reaser et al., 2007; Blackburn et al., 2014; Kumschick et al., 2015). Invasion can be viewed as a function of the relationship between abiotic characteristics of an invaded ecosystem, the traits of alien species and recipient community, in addition to propagule pressure that includes the number of non-native individuals in an introduced group and the number of introduction events (Lockwood et al., 2005; Catford et al., 2009). Each of these factors contribute to the success of an invasion, although their influence is unlikely to be equal, and in many cases, there are multiple, largely interactive factors underlying invasibility (Alpert et al., 2000; Catford et al., 2009, 2011).

Several hypotheses have been proposed for explaining successful invasions (Hierro et al., 2005; Jeschke, 2014). These hypotheses may actually be overlapping and they are not necessarily mutually exclusive. According to the "propagule pressure hypothesis" (H1), high supply and frequency of plant propagule introductions (including adult plants, seeds or reproductive vegetative fragments) should raise the probability of successful invasion through increased genetic diversity, seed swamping and ongoing addition of new individuals, greater probability of introduction into a favorable environment, and increasing the chances of overcoming environmental and demographic stochasticity (Lockwood et al., 2005, 2013; Catford et al., 2009; Simberloff, 2009). In the British Isles, the H1 may explain, to a degree, why many plant invasions occur close to human sites and settlements (Crawley et al., 1996).

The "biotic resistance hypothesis" predicts ecosystems will have greater resistance to invasion with high native species richness compared to ecosystems with low richness (Elton, 1958; Levine and D'Antonio, 1999; Lonsdale, 1999; Jeschke, 2014). It is a special formulation of the "diversity-stability hypothesis," a more general ecological hypothesis, according to which ecosystems with high biodiversity are more stable than ecosystems with low biodiversity. This is closely related to the "empty niche" hypothesis, which posits that when an ecosystem is unsaturated from low indigenous species diversity, invading groups will occupy underutilized niches and use available resources (MacArthur, 1970; Hierro et al., 2005). This hypothesis predicts a negative correlation between native and alien species richness. However, a number of studies reported the opposite pattern, i.e., a positive correlation between native and alien species richness (McKinney, 2002), leading to the formulation of the "acceptance hypothesis" and to a "rich-get-richer" pattern (Stohlgren et al., 2003, 2006). As "biotic resistance hypothesis vs. acceptance hypothesis" (H2) are essentially two sides of the same coin we have therefore treated them as one testable hypothesis, i.e., the chances of successful invasion depend (positively or negatively) on the native species richness.

The "disturbance-mediated hypothesis" (H3) states that disturbance can reset succession and improve availability of resources, providing alien species with an equal or improved chance of success in colonization and establishment (Catford et al., 2009). A strong global relationship has been proposed between human-induced disturbance and invasibility (Zhang et al., 2006), explained both by human activities creating pathways for the introduction of alien species and the ongoing generation of new (artificial or agricultural) habitat types due to human-mediated habitat destruction (of natural habitat types) and biodiversity loss.

According to the "environmental heterogeneity hypothesis" (H4), ecosystems can host a diversity of species when they contain a variety of niches and high environmental variability (Melbourne et al., 2007; Catford et al., 2009). Heterogeneity in key environmental factors, such as temperature, water, nutrients, sunlight and physical disturbances plays a definite role in community-level resistance to new invasions, as well as in their outcome and the severity of their impacts (Melbourne et al., 2007).

Islands are geographically isolated and often have a distinct combination of species compared to those found on continents, and are under particularly strong ecological threat from invasive alien species (Reaser et al., 2007). The Mediterranean islands are considered more vulnerable to invasive alien species introductions than their comparable closest mainland areas due to their proportionally lower native diversity and the presence of unsaturated local communities (Hulme et al., 2008), and evidence suggests there have been major increases in the number of invasive alien species recently (Celesti-Grapow et al., 2016). Moreover, the historical and modern importance of ports, trade and anthropogenic disturbance, with changes to land use and climate are likely facilitating further the introduction and establishment of alien species (Ross et al., 2008). Crete, the fifth largest Mediterranean island, and surrounding islets provide an interesting case study to test hypotheses on plant invasions, as the islands are a biodiversity hotspot containing a unique flora (Medail and Quezel, 1997). The islands have been populated by humans for over 9,000 years, but there are recent severe anthropogenic pressures and land use changes (Papanastasis and Kazaklis, 1998; Sarris et al., 2005). We sought to quantify the patterns of alien vascular plant species richness by testing four invasibility hypotheses.

METHODS

Study Area

The island of Crete and 49 surrounding islets (i.e., the Cretan area Turland et al., 1993) were examined, covering an area of 8,374 km^2 and including 1,647 vascular plant species (Turland et al., 1993). The Crete area exhibits rich plant diversity and is a red alert biodiversity hotspot within the Mediterranean (Médail and Diadema, 2009), and around 10% of its total native number of species comprises endemic vascular flora (Turland et al., 1993). Apart from the endemic species, the flora comprises significant Eastern Mediterranean and Anatolian elements, (e.g., *Datisca* sp), Tertiary relictual Aegean endemics (e.g., *Lomelosia* sp), and

North African species (*Erodium* sp). There are three mountain massifs in Crete, with Psiloritis the highest (at 2,456 m) and Lefka Ori the most extensive. The latter, situated at the western part on the island include 15 peaks above 2,200 m (Vogiatzakis et al., 2003). The distribution of elevations on the island changes steeply, particularly on its south-western part (Vogiatzakis et al., 2003). Coastal areas are highly modified mainly for tourism infrastructure, and the human population of the island mainly resides on the north side where most cities are located (and thus most human-created and modified habitats are present). There is a significant number of cultivated land areas (agricultural habitats) predominantly with olive trees and grapes.

Linking Hypothesis with Covariates

Statistical models were designed to explicitly test the four hypotheses. The covariates used to test H1 were the percentage

TABLE 1 | List of all models and predictor covariates of alien species richness related to each of the four hypotheses.

Hypothesis	Model	Predictor variables
	g0	Area
1	g1	Artificial coverage + area
	g2	Agricultural coverage + area
	g3	Artificial coverage + agricultural coverage + area
2	g4	Native species richness + area
	g5	Endemic species richness + area
	g6	Native species richness + endemic species richness + area
3	g7	Artificial richness + area
	g8	Agricultural richness + area
	g9	Artificial richness + agricultural richness + area
4	g10	Soil richness + area
	g11	Habitat richness + area
	g12	Altitude range + area
	g13	Temperature range + area
	g14	Precipitation range + area
	g15	Altitude range + habitat richness + area
	g16	Altitude range + soil richness + area
	g17	Altitude range + temperature range + area
	g18	Altitude range + precipitation range + area
	g19	Habitat richness + soil richness + area
	g20	Habitat richness + temperature range + area
	g21	Habitat richness + precipitation range + area
	g22	Soil richness + temperature range + area
	g23	Soil richness + precipitation range + area
	g24	Temperature range + precipitation range + area
	g25	Altitude range+ habitat richness + soil richness + area
	g26	Altitude range+ habitat richness + temperature range + area
	g27	Altitude range+ habitat richness + precipitation range + area
	g28	Altitude range+ soil richness + temperature range + area
	g29	Altitude range+ soil richness + precipitation range + area
	g30	Altitude range+ temperature range + precipitation range + area
	g31	Habitat richness+ soil richness + temperature range + area
	g32	Habitat richness+ soil richness + precipitation range + area
	g33	Habitat richness+ temperature range + precipitation range + area
	g34	Soil richness+ temperature range + precipitation range + area
	g35	Altitude range+ habitat richness+ soil richness + temperature range + area
	g36	Altitude range+ habitat richness+ soil richness + precipitation range + area
	g37	Habitat richness+ soil richness+ temperature range + precipitation range + area
	g38	Altitude range+ habitat richness+ soil richness+ temperature range + precipitation range + area

All models also included area as a covariate.

cover of artificial and agricultural habitats, with each providing a measure of human presence, activity and intensity of the main pathways for alien plants, i.e., ornamental planting, trade of contaminated commodities, and cultivation, (see European Alien Species Information Network; Katsanevakis et al., 2015), and used herein as a surrogate of propagule pressure. For H2, native and endemic species richness provides a measure of species diversity. For H3, artificial and agricultural habitat richness were used measuring the variety of anthropogenic habitats and thus available empty habitats deriving from human-mediated disturbance. Artificial plus agricultural percentage of cover was partly correlated with total habitat richness (linear regression, $R^2 = 23.1\%$) and thus to that end there is some co-linearity between the covariates for H1 and H3 (see Figure S1). For H4, environmental traits measuring heterogeneity were used; these included altitude range, habitat richness, soil richness, temperature range, and precipitation range. In the plant atlas by Turland et al. (1993) coastal cells have unequal surface area with inland cells. In order to account for this effect the total surface area of each cell was normalized by the surface area of inland cells ($8.25 \times 8.25 = 68.0625$ km^2). All hypotheses included the effects of cell land area as a fixed covariate in order to account for the effect of unequal land surface area of cells neighboring with the sea. Therefore, apart from cell area, all four hypotheses included different sets of covariates. A full list with all the statistical models explored and their link with the four hypotheses is provided in **Table 1**.

Plant Data

Maps of presence-absence vascular plant species distributions in the Cretan area were digitized from Turland et al. (1993) and its latest supplement (Chilton and Turland, 2008). The main island of Crete and the surrounding islands were divided into 162 grid cells of 8.25 × 8.25 km following the grid cell size of Turland et al. (1993). For each cell, the native, endemic, and alien species richness was calculated. We used (Turland et al., 1993; Chilton and Turland, 2008) and references therein to define native ($n_{nat} = 1{,}395$) and endemic ($n_{end} = 174$) species, and the vascular plants from D'Agata et al. (2009) that are listed in Turland et al. (1993) and Chilton and Turland (2008) were used to define alien ($n_{alien} = 78$) species.

Land Cover Data

Habitat classification relied on the most detailed resolution available of the CORINE Landcover (level 3, spatial resolution 100 m; EEA, 2010), to calculate the richness and percentage of every land cover class within every grid cell, using Patch Analyst 5.1 within ArcGIS. In order to avoid potential temporal deviance between habitat classifications and species presences in cells, the last updated available supplement for the flora of Crete published in 2008 (Chilton and Turland, 2008) and the closest available time snapshot of the CORINE landcover for Crete in 2010 were used. The classification process resulted in 29 habitat types, of which 9 were agricultural, 7 were artificial, and 13 were natural. We recorded habitat richness per cell as the number of different land cover types present on each cell as well as their percentage of cover.

Climatic, Soil, and Altitude Data

Climatic variables were derived from WorldClim (Hijmans et al., 2005) for Crete and surrounding islets. The original resolution of the climatic data was 1 km. In order to re-scale them to 8.25 km and match them with the grid of the plant data the mean values of the 1 km data within the 8.25 km cells were calculated and used. The climatic variables used here were: annual mean temperature (*Tempmean*), annual mean temperature of warmest quarter (*Tempwarm*), annual mean temperature of coldest quarter (*Tempcold*) all in °C, annual mean precipitation (*Precipmean*), precipitation of wettest quarter (*Precipwet*), precipitation of driest quarter (*Precipdry*) all in mm year^{-1}. Soil data were derived from SoilGrid (Hengl et al., 2014) and were rescaled from 1 to 8.25 km as done for the climatic data. The soil variable used was soil richness in the cell (*Soildiv*) derived as the number of different soil types occurring within each cell. The indices of elevation recorded were the mean of all elevation scores within the cell (Alt) and the range of elevation within the cell (*Alt range*) both in meters.

Statistical Analysis

The inherently non-linear relationship between alien species richness and potential explanatory covariates were modeled with Generalized Additive Models (GAMs); (Hastie and Tibshirani, 1990). GAMs are a generalization of the multiple regression model maintaining the additive nature of the model replacing the terms of the linear equation with non-parametric smooth functions $s(X_i)$ of each independent variable (X_i). In GAMs the distribution of the dependent variable can be explicitly quantified and therefore does not need to be normal nor does it have to be continuous. Having specified the underlying distribution of the dependent variable, values are predicted from a linear combination of independent variables, via a link function which specifies the distribution of the dependent variable:

$$g\left(E\left[Y\right]\right) = b_0 + s_1\left(X_1\right) + \ldots + s_m(X_m)$$

where Y is the dependent variable, $E[Y]$ is the expected value of Y and $g(E[Y])$ is the link function, that links the dependent variable to the predictor variables X_i.

Specifically, the expected value of alien species richness in each 8.25 × 8.25 km cell was related to the explanatory covariates, according to the general formulation:

$$g\left(E\left[Richness_i\right]\right) = c + \sum_m s_m(X_{mi})$$

where g is the link function, $s_m(\cdot)$ is the one-dimensional smooth function for covariate m, and X_{mi} is the value of covariate X_m for the ith cell. All the smooth functions $s_m(\cdot)$ were implemented using penalized cubic regression splines (with the default basis dimension $q = 10$, estimated by penalized iterative least squares; Wood, 2006). Optimum smoothing was determined using generalized cross validation (GCV), increasing by a factor of $\gamma = 1.4$ in the GCV score the amount that the effective degree of freedom of each model counts. As occasionally the GCV method may result in over-fitting a value of $\gamma \approx 1.4$ was

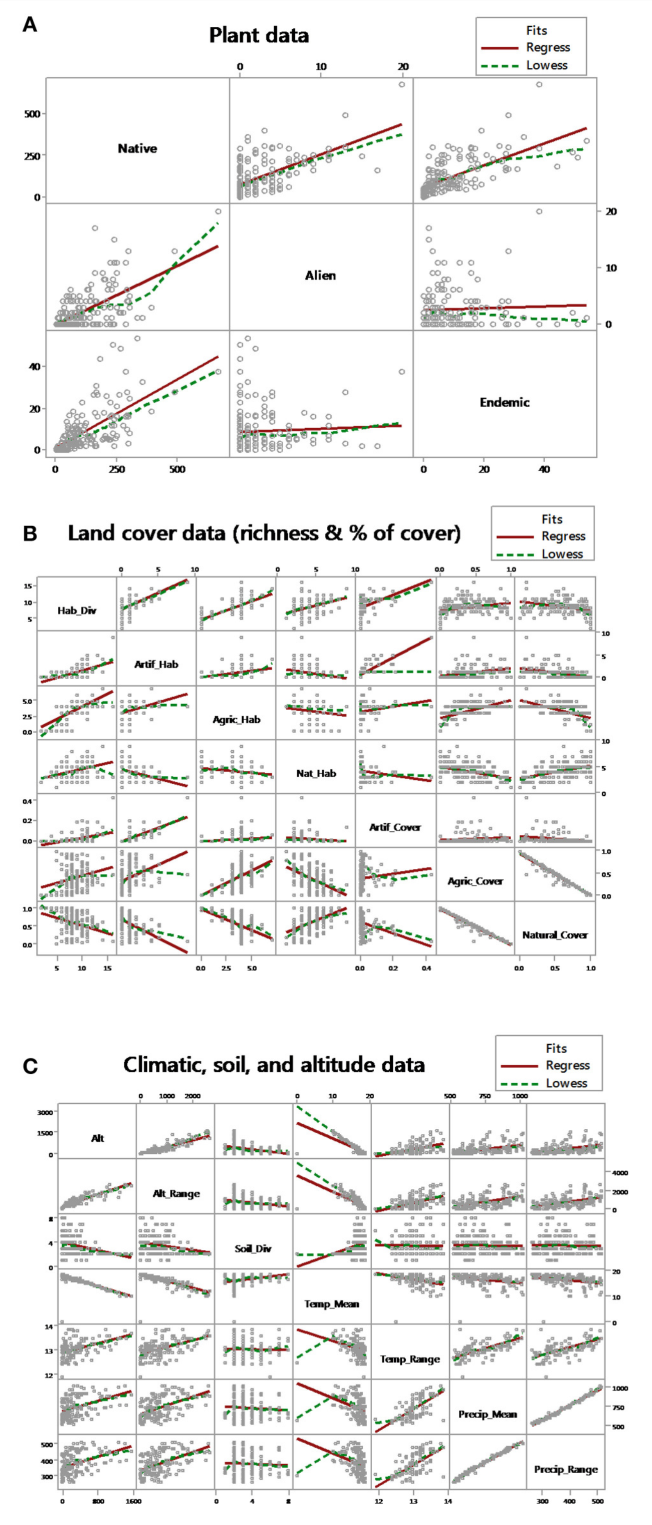


FIGURE 1 | Initial data mining: **(A)** exploring the relationship between species richness data (described in the main text, methods "Plant data") **(B)** land cover indices (described in the main text, methods "Land cover data"), and **(C)** environmental and climatic data (described in the main text, methods "Climatic, soil, and altitude data") with multi-scatter plots. Grey circles indicate data points, solid red lines linear regression, and dashed green lines a lowess smoother between each two data indices.

used in order to account for this without compromising model fit (Kim and Gu, 2004). The model fitting was conducted using the mgcv package (Wood, 2006) in R v.3.1.2 (R Development Core Team, 2016).

To identify the best underlying probability distribution, five different error distributions were examined including Gaussian, Lognormal, Gamma, Poisson, and Negative binomial; these distributions were compared based on a full model with all of the predictor variables included using the Akaike Information Criterion (Akaike, 1973; Burnham and Anderson, 2002). After identifying the best probability distribution, AIC was used for model selection among the set of candidate models. Thirty-nine alternative GAMs g_i, $i = 0\text{–}38$ were fitted, where a reference model g_0 was used with area as the only predictor variable. The subsequent 39 models all involved one or more predictor variables related with the four hypotheses (see **Table 1** for the full list). The equations of the maximal model for H1, H2, H3, and H4, respectively are:

$$
\begin{aligned}
\text{H1}: g_3\left(E\left[Alien\right]\right) &= s_1\left(Artif.\ cover\right) + s_2\left(Agric.\ cover\right) \\
&\quad + s_3(area) \\
\text{H2}: g_6\left(E\left[Alien\right]\right) &= s_1\left(Native\ spp\ richness\right) \\
&\quad + s_2\left(Endemic\ spp\ richness\right) \\
&\quad + s_3(area) \\
\text{H3}: g_9\left(E\left[Alien\right]\right) &= s_1\left(Artif.\ richness\right) \\
&\quad + s_2\left(Agric.\ richness\right) + s_3(area) \\
\text{H4}: g_{38}\left(E\left[Alien\right]\right) &= s_1\left(Alt.\ range\right) + s_2\left(Habitat\ richness\right) \\
&\quad + s_3\left(Soil\ richness\right) + s_4\left(Temp.\ range\right) \\
&\quad + s_5\left(Precip.\ range\right) + s_7(area)
\end{aligned}
$$

The model with the lowest AIC was selected as the best model. An empirical semivariogram of the residuals of the optimal model showed no evidence of spatial autocorrelation (Legendre and Fortin, 1989).

Variance partitioning was employed in order to provide a further measure of how much independent variation is accounted for by each of the four hypotheses (Borcard and Legendre, 1994). We added together all the relevant covariates to each hypothesis and quantified the independent contribution of each hypothesis to the total pattern of alien species richness (H1 included variables as described in the model g_3, H2 as in g_6, H3 as in g_9, and H4 as in g_{38}). Randomizations ($n = 99$) were also performed in order to asses confidence intervals (Mac Nally, 2002).

RESULTS

An initial data mining exploration between all dependent and independent variables with multi-scatter plots is provided in **Figure 1** and a cluster analysis is provided in the Supplementary Material (Figure S1). From a simple initial linear regression or smoothing perspective, alien species richness have a positive relationship with native species richness, and no relationship (regression slope close to zero) or weakly negative based on smoothing trend lines with endemic species richness

(**Figure 1A**). Endemic species richness was positively correlated with native species richness (**Figure 1A**). Overall the relationship between richness and cover is not straightforward as either regression slopes are relatively low or there is a deviance between regressions and smoothing (**Figure 1B**). Altitude is positively correlated with altitude range and mean precipitation with precipitation range but temperature mean has no clear relationship with temperature range (**Figure 1C**). As expected altitude is negatively correlated with mean temperature, but not with temperature range (**Figure 1C**). Altitude is also negatively correlated with soil richness (**Figure 1C**).

The Poisson probability distribution had the lowest AIC score (Table S1), and thus all models were run using a Poisson error distribution and log as the natural link function. From the full analyses of all 39 models (Table S2), the model g_6 related with the H2 was the best-supported model with the lowest AIC (550.7793) and explained 67.8% of deviance. Model g_6 included native species richness, endemic species richness and area as predictor variables. Alien species richness was positively correlated with native species richness, negatively correlated with endemic species richness, and, as expected, increased with area across the main range of cell land areas (the slightly declining richness with area at high values of area was non-significant and was based on a very low number of data points); (**Figure 2**).

Variance partitioning of the habitat and environmental covariates included in each hypothesis showed the variables involved in H2 explained close to 60% of variation in alien richness, whereas H1 explained just over 10%, H3 close to 30%, and H4 under 10% of variation (**Figure 3**).

DISCUSSION

Patterns of alien species in the Cretan area, based on testable hypotheses, provided the greatest support for the "biotic

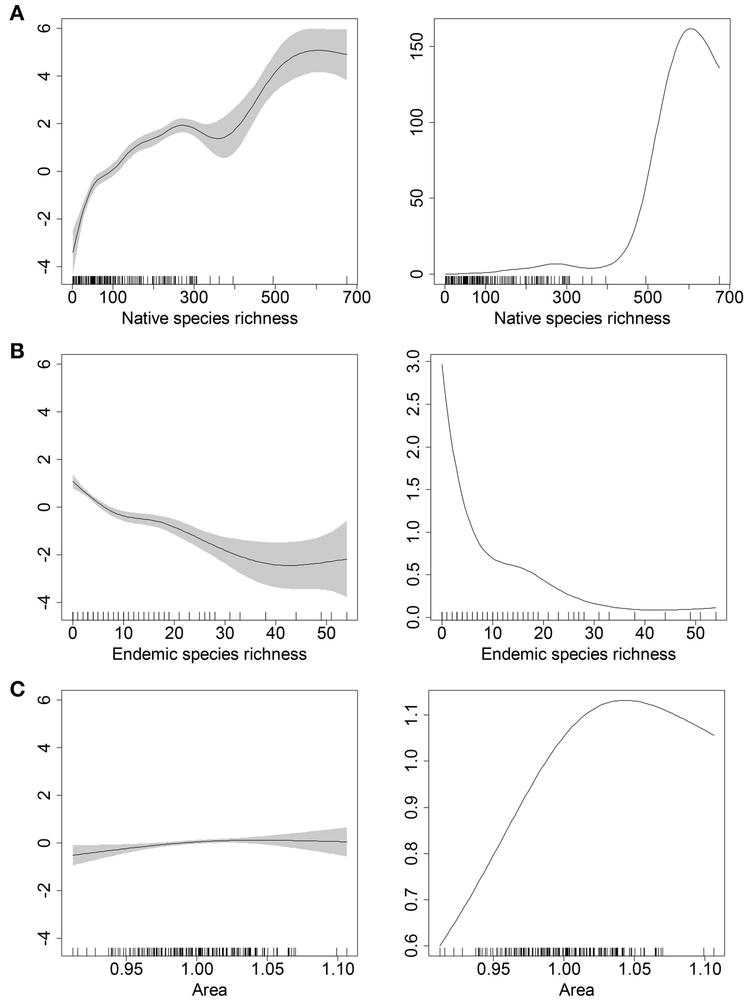


FIGURE 2 | Estimated smooth terms for the best model of alien species richness. The top panels **(A)** show the positive correlation between alien species richness and native species richness, **(B)** the middle panel shows the negative correlation with endemic species richness, and **(C)** bottom panel the relationship with area. For panels on the left, smooth terms (black lines) are given in the linear predictor scale and the corresponding 95% confidence intervals are indicated with gery intervals. For panels on the right, smooth terms are given in the response scale of the raw data (i.e., smooth terms were exp-transformed). The numbers after the commas in the y-axis titles are the estimated degrees of freedom of the smooth terms. Rug on the horizontal axis represents density of datapoints.

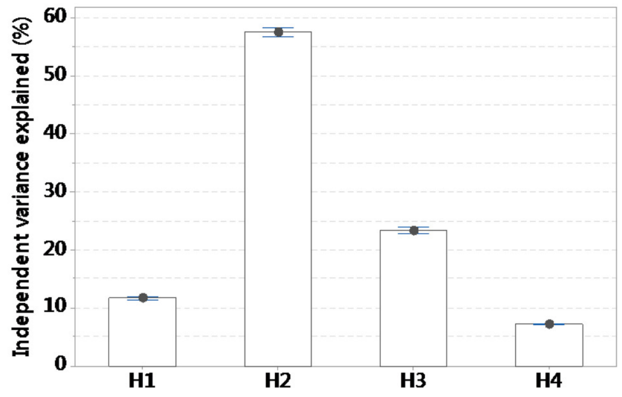


FIGURE 3 | Variance partitioning of environmental and habitat covariates included in each of four hypotheses. H1, propagule pressure hypothesis; H2, biotic resistance vs. acceptance hypotheses; H3, disturbance-mediated hypothesis; H4, environmental heterogeneity hypothesis.

resistance vs. acceptance hypotheses" (H2) both in terms of model selection out of 39 models as well as from variance partitioning—nearly 60% of the total variance was explained by H2. This does not mean there is no relationship between alien species richness and human activity and environmental disturbance. Indeed, other models than the optimal had strong explanatory power and all hypotheses showed patterns as expected with alien species richness been positively correlated with propagule pressure (Wallace and Prather, 2016), human activity (Malavasi et al., 2016), and environmental disturbance (Peltzer et al., 2016) as well as synergistic interactions (Barney et al., 2016; Berg et al., 2016). These results are only partly shown here, as the deviance explained by each model and variance partitioning show the magnitude of the result but not the sign of the relationship, and interactions of covariates between different hypotheses were not explored.

Our results indicate that rather than a "black and white" or a "single variable" attempt to explain patterns of alien species, the reality seems to lie in between—all hypotheses contributed to explaining the pattern. The concept of a single "true" predictive model seems to be of little utility in ecology (Evans et al., 2014) with more data resulting in more complex models (Evans and Moustakas, 2016). Nevertheless, in this study the goal was not to develop the best possible predictive model (i.e., explaining as much of the variance as possible) but rather to investigate the differential support of the main hypotheses in invasion biology. Note that several more testable hypotheses exist—(see e.g., Hierro et al., 2005; Catford et al., 2009; Jeschke, 2014)—regarding invasibility that we did not explore because some of them are overlapping and would require using the same covariates to explore different hypotheses.

According to the results derived here there is a strong implication for the acceptance hypothesis and a "rich-get-richer" distribution of invasive vascular plant species in the Cretan area, whereby regions with high native species richness are associated with high alien species richness (Stohlgren et al., 2003). The positive correlation between alien and native species richness is consistent with multiple previous large-scale observational studies (Knops et al., 1995; Robinson et al., 1995; Planty-Tabacchi et al., 1996; Levine and D'Antonio, 1999; Lonsdale, 1999; Levine, 2000; McKinney, 2002; Stohlgren et al., 2003, 2006; Souza et al., 2011; Blackburn et al., 2016). It is difficult to ascertain the full causal relationship between alien and native species distributions, as their apparent co-variation may be due to similar, independent reactions to the same factors, specifically favorable environmental conditions, adequate resources and fertile land, meaning the conditions good for native species are also good for alien species (Levine and D'Antonio, 1999; Stohlgren et al., 1999; Richardson and Pyšek, 2006).

There are several studies reporting an inverse relationship between alien and native species, in terms e.g., of environmental energy and human impact (Marini et al., 2009), responses to soil fertility and in general of trait or performance differences among plant functional groups (Peltzer et al., 2016). However, the existence of different patterns between alien and native species is not consistent worldwide (Ordonez and Olff, 2013), and the result of inverse patterns between alien and endemic species reported here appears to be fairly novel.

We performed a *post-hoc* analysis examining separately the relationship of alien and endemic species richness with altitude and habitat richness, and the correlation between endemic and native species richness (**Figure 4**). There is a positive correlation between endemic species richness and altitude, the opposite pattern found for alien species (**Figures 4A,B**), indicating the observed relationship between alien and endemic species richness is possibly linked with altitude. Indeed, almost half (43%) of Cretan endemics are found solely at elevations higher than 1,000 m (Legakis and Kypriotakis, 1994). Endemic species richness is known to increase with altitude (Cañadas et al., 2014), and typically peaks at higher altitudes compared to total species richness (Vetaas and Grytnes, 2002). Alien and endemic species distributions have been shown to differ along a longitudinal gradient in rivers in another Mediterranean island (Sardinia), with endemic species more common in natural upper parts and alien species more common in lower and middle parts, the latter linked to human impact and agriculture (Angiolini et al., 2013). In Tenerife (Canary islands) altitude was the most important factor determining (inverse) patterns in endemic and native species along roadsides (Bacaro et al., 2015). Altitude itself is likely to be masking several long-term acting processes regarding endemism: the increased proportion of Cretan endemics with altitude is regarded as being due to diversification linked to the Cretan mountains (especially mid-altitude areas), driven by altitudinal-driven ecological isolation (Trigas et al., 2013). Climate filtering reflected in varying patterns along altitudinal gradients is an important determinant of the richness of alien species (which are not adapted to high altitudes), while anthropogenic pressures may explain the richness of alien species at low elevation (Bacaro et al., 2015). Higher elevations have an interaction effect between area and habitat diversity: area since in small islands land surface area extends with elevation, and components of habitat diversity, such as more heterogeneous topography, a broader selection of micro-habitats, and a reduced effect of the sea are related with elevation

(Panitsa et al., 2006). It is therefore possible that the effect of environmental heterogeneity is hidden behind elevation.

Our *post-hoc* analysis indicated a further variable of interest, habitat richness, which showed a positive correlation with alien species richness (**Figure 4C**), supporting the H4 but no consistent effect on endemic species (**Figure 4D**). This shows the two groups have different responses to habitat richness, but the inconclusive results for endemic species seem to rule out habitat richness as a primary driver of the relationship between alien and endemic species richness. In this case this implies that habitat richness is explored differently by alien and endemics potentially because alien species are habitat generalists while endemics are specialists

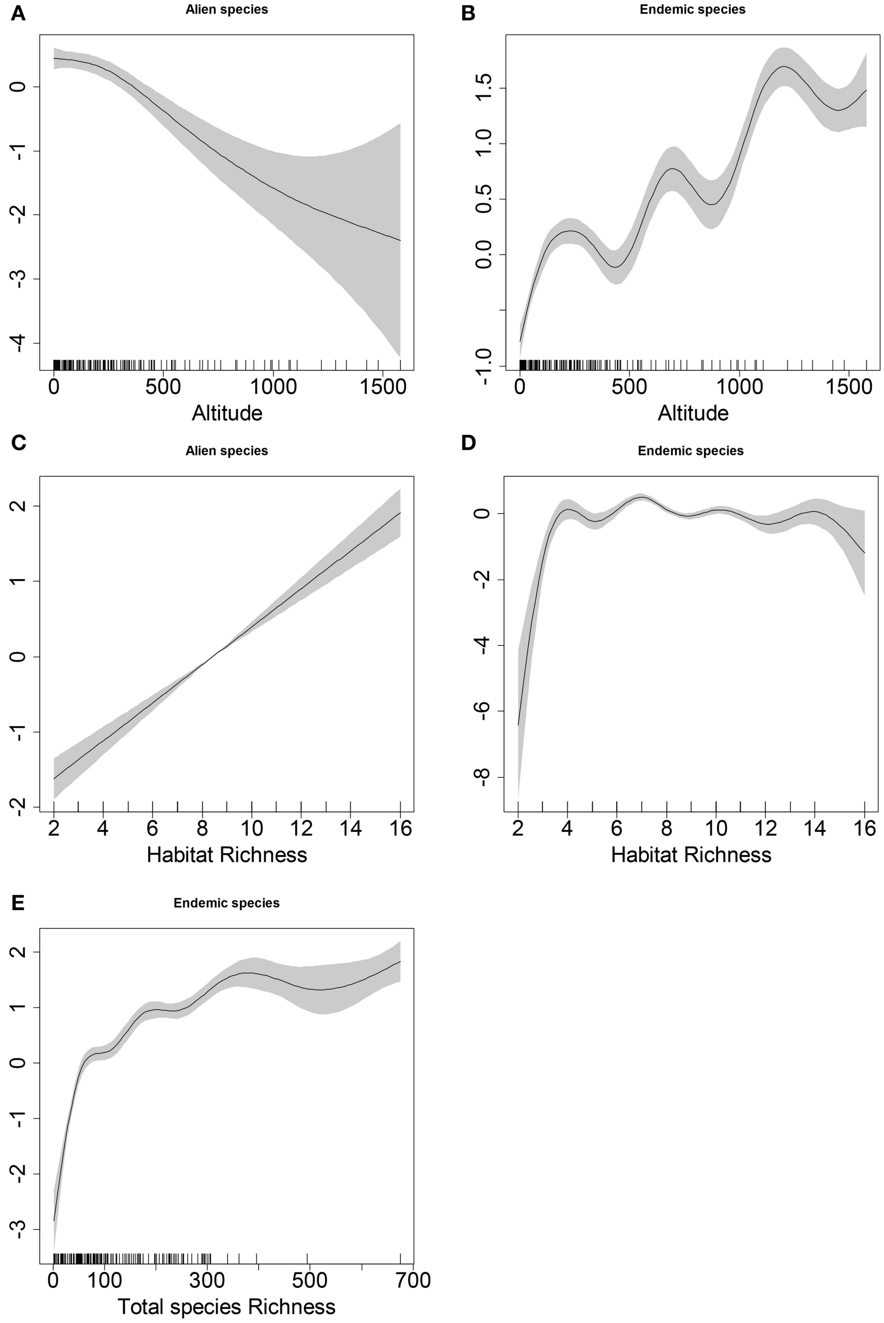


FIGURE 4 | Estimated smooth terms for **(A)** alien species richness and altitude, **(B)** endemic species richness and altitude, **(C)** alien species richness and habitat richness, **(D)** endemic species richness and habitat richness, and **(E)** endemic species richness and native species richness. Smooth terms (black lines) are given in the linear predictor scale and the corresponding 95% confidence intervals are indicated with gray intervals. The numbers after the commas in the y-axis titles are the estimated degrees of freedom of the smooth terms. Rug on the horizontal axis represents density of datapoints.

(Harrison, 1999; Evangelista et al., 2008). More often than not, high habitat richness locations contain agricultural or artificial habitats (i.e., high habitat richness is associated with non-natural, modified habitats). This is particularly important as due to climatic change, disproportionately more high altitude habitats are likely to be modified (Louca et al., 2015), thereby increasing habitat richness in high altitute locations (Dirnböck et al., 2011). Additional *post-hoc* analysis regarding alien species richness exhibited no consistent trends against indices of temperature (mean, warmest, and coldest) or precipitation (mean, wettest, and driest); (Figure S2).

Understanding the mechanisms by which alien plants impact native (Vilà et al., 2011) or endemic species (Thomson, 2005) is a key question for understanding alien species dynamics. The interaction between alien and native species may also be positive through mutualism or facilitation (Mooney and Cleland, 2001; Rodriguez, 2006). A key question that arises here is whether patterns of endemic species can provide long-term resistance indicators to invasions, or if the inverse pattern between alien and endemics is simply a result of isolation and inaccessibility due to high altitude. Eitherway this has strong implications for conservation: high altitude locations would need to remain unmodified as their habitat modification could result in increased habitat richness that in turn can be explored by alien species (Marvier et al., 2004). While speciation in endemic species could be linked to their biogeographical and genetic isolation from closely related species (Cañadas et al., 2014), the positive correlation with native species richness shows more endemic species are present in areas with higher richness (**Figure 4E**), and the dynamics of the relationship between endemic and alien species richness are not simply proxies for relationships with total species richness. In addition, endemic and alien components of the flora are not random assemblages of species: there are more alien than endemic families, and the pool of alien species is less of a random selection of species than it is for endemics (Vilà and López-Darias, 2006).

CONCLUSIONS

Using the Mediterranean island of Crete as a case study, we suggest a framework for analyzing spatial data of alien species distributions, based on environmental predictors, aiming to gain an understanding of their spatial patterns and spread. The analysis involved a three-way coupling of data [(i) spatial plant species distributions, (ii) land-use data, and (iii) physical geography-related data, such as climatic, altitudinal, and soil data]. Technological developments in smart sensors, social networks, and digital maps, spatio-temporal data are more available than ever before (Reis et al., 2015; Miyazaki et al., 2016; Niphadkar and Nagendra, 2016) and ecology in the big data era needs to integrate novel methods for their analysis (Moustakas, 2017). The availability of large datasets poses great challenges in data analytics (Moustakas and Evans, 2017) but also increased availability of computing power facilitates the use of computationally-intensive methods for the analysis of such data in ecology (Moustakas and Evans, 2015).

Despite the relatively large volume of data, the analysis conducted was on a hypothesis-based model selection rather than statistical-only model selection. Recent debates in data analytics have argued that the availability of a large volume of data will make the scientific method obsolete (Anderson, 2008); hypothesis-driven, or equation-driven research will become irrelevant and hypothesis-free data mining will be used instead (Anderson, 2008). This analysis showcases that if one simply relies on data-driven science several components of scientific methods could be made poorer (Moustakas, 2017). This study has identified an inverse relation between alien and endemic species richness, to our knowledge a novel result. The result was derived via a correlative analysis which does not show causality nor provide a mechanism. However, it has identified the potential for learning patterns, traits, and proccesses from endemic species as counter preferences or potentially competitive exclusion between alien and endemic species. This could be achieved via (a) field experiments and mechanistic studies (b) biogeographic analysis of alien and endemics and (c) exploring additional variables than the ones explored here at a macroecological scale as potential traits.

AUTHOR CONTRIBUTIONS

AB performed the majority of the statistical analysis, and wrote a significant part of the manuscript. SK formulated the hypotheses, contributed to the statistical analysis and to manuscript writing. IV performed the land use, soil, climatic, and physical geography remote sensing analysis resulting into the environmental dataset, and commented on the manuscript. AG provided the total, endemic and alien species data and commented on the manuscript. AM supervised and coordinated the research, performed part of the statistical analysis, and wrote a significant part of the manuscript.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: http://journal.frontiersin.org/article/10.3389/fevo.2017.00091/full#supplementary-material

REFERENCES

Akaike, H. (1973). Maximum likelihood identification of Gaussian autoregressive moving average models. *Biometrika* 60, 255–265. doi: 10.1093/biomet/60.2.255

Alpert, P., Bone, E., and Holzapfel, C. (2000). Invasiveness, invasibility and the role of environmental stress in the spread of non-native plants. *Perspect. Plant Ecol. Evol. Syst.* 3, 52–66. doi: 10.1078/1433-8319-00004

Anderson, C. (2008, July 16). The end of theory. *Wired Magazine* 16.

Angiolini, C., Nucci, A., Landi, M., and Bacchetta, G. (2013). Distribution of endemic and alien plants along Mediterranean rivers: a useful tool to identify areas in need of protection? *C. R. Biol.* 336, 416–423. doi: 10.1016/j.crvi.2013.07.001

Bacaro, G., Maccherini, S., Chiarucci, A., Jentsch, D., Rocchini, D., Torri, M., et al. (2015). Distributional patterns of endemic, native and alien species along a roadside elevation gradient in Tenerife, Canary Islands. *Commun. Ecol.* 16, 223–234. doi: 10.1556/168.2015.16.2.10

Barney, J. N., Ho, M. W., and Atwater, D. Z. (2016). Propagule pressure cannot always overcome biotic resistance: the role of density-dependent establishment in four invasive species. *Weed Res.* 56, 208–218. doi: 10.1111/wre.12204

Berg, J. A., Meyer, G. A., and Young, E. B. (2016). Propagule pressure and environmental conditions interact to determine establishment success of an invasive plant species, glossy buckthorn (*Frangula alnus*), across five different wetland habitat types. *Biol. Invasions* 18, 1363–1373. doi: 10.1007/s10530-016-1073-x

Blackburn, T. M., Delean, S., Pyšek, P., and Cassey, P. (2016). On the island biogeography of aliens: a global analysis of the richness of plant and bird species on oceanic islands. *Global Ecol. Biogeogr.* 25, 859–868. doi: 10.1111/geb.12339

Blackburn, T. M., Essl, F., Evans, T., Hulme, P. E., Jeschke, J. M., Kühn, I., et al. (2014). A unified classification of alien species based on the magnitude of their environmental impacts. *PLoS Biol.* 12:e1001850. doi: 10.1371/journal.pbio.1001850

Borcard, D., and Legendre, P. (1994). Environmental control and spatial structure in ecological communities: an example using oribatid mites (Acari, Oribatei). *Environ. Ecol. Stat.* 1, 37–61. doi: 10.1007/BF00714196

Burnham, K. P., and Anderson, D. R. (2002). *Model Selection and Multimodel Inference*. New York, NY: Springer Verlag.

Cañadas, E. M., Fenu, G., Peñas, J., Lorite, J., Mattana, E., and Bacchetta, G. (2014). Hotspots within hotspots: endemic plant richness, environmental drivers, and implications for conservation. *Biol. Conserv.* 170, 282–291. doi: 10.1016/j.biocon.2013.12.007

Catford, J. A., Jansson, R., and Nilsson, C. (2009). Reducing redundancy in invasion ecology by integrating hypotheses into a single theoretical framework. *Divers. Distrib.* 15, 22–40. doi: 10.1111/j.1472-4642.2008.00521.x

Catford, J. A., Vesk, P. A., White, M. D., and Wintle, B. A. (2011). Hotspots of plant invasion predicted by propagule pressure and ecosystem characteristics. *Divers. Distrib.* 17, 1099–1110. doi: 10.1111/j.1472-4642.2011.00794.x

Celesti-Grapow, L., Bassi, L., Brundu, G., and D'Auria, U. (2016). Plant invasions on small Mediterranean islands: an overview. *Plant Biosys. Int. J. Dealing Aspects Plant Biol.* 150, 1119–1133. doi: 10.1080/11263504.2016.1218974

Chilton, L., and Turland, N. J. (2008). *Flora of Crete a Supplement II, Additions 1997-2008*. Available online at: http://www.marengowalkscom/fcshtml

Crawley, M. J., Harvey, P. H., and Purvis, A. (1996). Comparative ecology of the native and Alien Floras of the British Isles. *Philos. Trans.* 351, 1251–1259. doi: 10.1098/rstb.1996.0108

D'Agata, C. D. C., Skoula, M., and Brundu, G. (2009). A preliminary inventory of the alien flora of Crete (Greece). *Bocconea* 23, 301–315. Available online at: http://www.herbmedit.org/bocconea/23.pdf#page=301

Dirnböck, T., Essl, F., and Rabitsch, W. (2011). Disproportional risk for habitat loss of high-altitude endemic species under climate change. *Glob. Change Biol.* 17, 990–996. doi: 10.1111/j.1365-2486.2010.02266.x

EEA (2010). *CORINE Land Cover 2010*. Copenhagen: European Environment Agency.

Elton, C. S. (1958). *The Ecology of Invasions by Animals and Plants*. London: Methuen.

Evangelista, P. H., Kumar, S., Stohlgren, T. J., Jarnevich, C. S., Crall, A. W., Norman, J. B. III., et al. (2008). Modelling invasion for a habitat generalist and a specialist plant species. *Divers. Distrib.* 14, 808–817. doi: 10.1111/j.1472-4642.2008.00486.x

Evans, M. R., Benton, T. G., Grimm, V., Lessells, C. M., O'Malley, M. A., Moustakas, A., et al. (2014). Data availability and model complexity, generality, and utility: a reply to Lonergan. *Trends Ecol. Evol.* 29, 302–303. doi: 10.1016/j.tree.2014.03.004

Evans, M. R., and Moustakas, A. (2016). A comparison between data requirements and availability for calibrating predictive ecological models for lowland UK woodlands: learning new tricks from old trees. *Ecol. Evol.* 6, 4812–4822. doi: 10.1002/ece3.2217

Harrison, S. (1999). Local and regional diversity in a patchy landscape: native, alien, and endemic herbs on serpentine. *Ecology* 80, 70–80. doi: 10.1890/0012-9658(1999)080[0070:LARDIA]2.0.CO;2

Hastie, T., and Tibshirani, R. (1990). *Generalized Additive Models*. New York, NY: Chapman&Hall/CRC.

Hengl, T., de Jesus, J. M., MacMillan, R. A., Batjes, N. H., Heuvelink, G. B. M., Ribeiro, E., et al. (2014). SoilGrids1km—global soil information based on automated mapping. *PLoS ONE* 9:e105992. doi: 10.1371/journal.pone.0105992

Hierro, J. L., Maron, J. L., and Callaway, R. M. (2005). A biogeographical approach to plant invasions: the importance of studying exotics in their introduced and native range. *J. Ecol.* 93, 5–15. doi: 10.1111/j.0022-0477.2004.00953.x

Hijmans, R. J., Cameron, S. E., Parra, J. L., Jones, P. G. and Jarvis, A. (2005). Very high resolution interpolated climate surfaces for global land areas. *Int. J. Climatol.* 25, 1965–1978. doi: 10.1002/joc.1276

Hulme, P. E., Brundu, G., Camarda, I., Dalias, P., Lambdon, P., Lloret, F., et al. (2008). "Assessing the risks to Mediterranean islands ecosystems from alien plant introductions," in *Plant Invasions: Human Perception, Ecological Impacts and Management,* eds B. Tokarska-Guzik, J. H. Brock, G. Brundu, L. Child, C. C. Daehler, P. Pyšek (Leiden: Backhuys Publishers), 39–56.

Jeschke, J. M. (2014). General hypotheses in invasion ecology. *Divers. Distrib.* 20, 1229–1234. doi: 10.1111/ddi.12258

Katsanevakis, S., Deriu, I., D'Amico, F., Nunes, A. L., Sanchez, S. P., Crocetta, F., et al. (2015). European alien species information network (EASIN): supporting European policies and scientific research. *Manag. Biol. Invasions* 6, 147–157. doi: 10.3391/mbi.2015.6.2.05

Kim, Y., and Gu, C. (2004). Smoothing spline Gaussian regression: more scalable computation via efficient approximation. *J. R. Stat. Soc. B* 66, 337–356. doi: 10.1046/j.1369-7412.2003.05316.x

Knops, J. M. H., Griffin, J. R., and Royalty, A. C. (1995). Introduced and native plants of the Hastings Reservation, central coastal California: a comparison. *Biol. Conserv.* 71, 115–123. doi: 10.1016/0006-3207(94)00008-E

Kolar, C. S., and Lodge, D. M. (2001). Progress in invasion biology: predicting invaders. *Trends Ecol. Evol.* 16, 199–204. doi: 10.1016/S0169-5347(01)02101-2

Kumschick, S., Gaertner, M., Vilà, M., Essl, F., Jeschke, J. M., Pyšek, P., et al. (2015). Ecological impacts of alien species: quantification, scope, caveats, and recommendations. *Bioscience* 65, 55–63. doi: 10.1093/biosci/biu193

Legakis, A., and Kypriotakis, Z. (1994). A biogeographical analysis of the Island of Crete, Greece. *J. Biogeogr.* 21, 441–445. doi: 10.2307/2845761

Legendre, P., and Fortin, M. J. (1989). Spatial pattern and ecological analysis. *Vegetatio* 80, 107–138. doi: 10.1007/BF00048036

Levine, J. M. (2000). Species diversity and biological invasions: relating local process to community pattern. *Science* 288, 852–854. doi: 10.1126/science.288.5467.852

Levine, J. M., and D'Antonio, C. M. (1999). Elton revisited: a review of evidence linking diversity and invasibility. *Oikos* 87, 15–26. doi: 10.2307/3546992

Lockwood, J. L., Cassey, P., and Blackburn, T. (2005). The role of propagule pressure in explaining species invasions. *Trends Ecol. Evol.* 20, 223–228. doi: 10.1016/j.tree.2005.02.004

Lockwood, J. L., Hoopes, M. F., and Marchetti, M. P. (2013). *Invasion Ecology, 2nd Edn.* Malden, MA: Wiley-Blackwell.

Lonsdale, W. M. (1999). Global patterns of plant invasions and the concept of invasibility. *Ecology* 80, 1522–1536. doi: 10.1890/0012-9658(1999)080[1522:GPOPIA]2.0.CO;2

Louca, M., Vogiatzakis, I. N., and Moustakas, A. (2015). Modelling the combined effects of land use and climatic changes: coupling bioclimatic modelling with Markov-chain Cellular Automata in a case study in Cyprus. *Ecol. Inform.* 30, 241–249. doi: 10.1016/j.ecoinf.2015.05.008

MacArthur, R. (1970). Species packing and competitive equilibrium for many species. *Theor. Popul. Biol.* 1, 1–11. doi: 10.1016/0040-5809(70)90039-0

Mac Nally, R. (2002). Multiple regression and inference in ecology and conservation biology: further comments on identifying important predictor variables. *Biodivers. Conserv.* 11, 1397–1401. doi: 10.1023/A:1016250716679

Malavasi, M., Santoro, R., Cutini, M., Acosta, A. T. R., and Laura Carranza, M. (2016). The impact of human pressure on landscape patterns and plant species richness in Mediterranean coastal dunes. *Plant Biosyst.* 150, 73–82. doi: 10.1080/11263504.2014.913730

Marini, L., Gaston, K. J., Prosser, F., and Hulme, P. E. (2009). Contrasting response of native and alien plant species richness to environmental energy and human impact along alpine elevation gradients. *Global Ecol. Biogeogr.* 18, 652–661. doi: 10.1111/j.1466-8238.2009.00484.x

Marvier, M., Kareiva, P., and Neubert, M. G. (2004). Habitat destruction, fragmentation, and disturbance promote invasion by habitat generalists in a multispecies metapopulation. *Risk Anal.* 24, 869–878. doi: 10.1111/j.0272-4332.2004.00485.x

McKinney, M. L. (2002). Influence of settlement time, human population, park shape and age, visitation and roads on the number of alien plant species in protected areas in the USA. *Divers. Distribut.* 8, 311–318. doi: 10.1046/j.1472-4642.2002.00153.x

Médail, F., and Diadema, K. (2009). Glacial refugia influence plant diversity patterns in the Mediterranean Basin. *J. Biogeogr.* 36, 1333–1345. doi: 10.1111/j.1365-2699.2008.02051.x

Medail, F., and Quezel, P. (1997). Hot-spots analysis for conservation of plant biodiversity in the Mediterranean Basin. *Ann. Missouri Bot. Gard.* 84, 112–127. doi: 10.2307/2399957

Melbourne, B. A., Cornell, H. V., Davies, K. F., Dugaw, C. J., Elmendorf, S., Freestone, A. L., et al. (2007). Invasion in a heterogeneous world: resistance, coexistence or hostile takeover? *Ecol. Lett.* 10, 77–94. doi: 10.1111/j.1461-0248.2006.00987.x

Miyazaki, Y., Teramura, A., and Senou, H. (2016). Biodiversity data mining from Argus-eyed citizens: the first illegal introduction record of Lepomis macrochirus macrochirus Rafinesque, 1819 in Japan based on Twitter information. *ZooKeys* 569, 123–133. doi: 10.3897/zookeys.569.7577

Mooney, H. A., and Cleland, E. E. (2001). The evolutionary impact of invasive species. *Proc. Natl. Acad. Sci. U.S.A.* 98, 5446–5451. doi: 10.1073/pnas.091093398

Moustakas, A. (2017). Spatio-temporal data mining in ecological and veterinary epidemiology. *Stoch. Environ. Res. Risk Assess.* 31, 829–834. doi: 10.1007/s00477-016-1374-8

Moustakas, A., and Evans, M. (2015). Coupling models of cattle and farms with models of badgers for predicting the dynamics of bovine tuberculosis (TB). *Stoch. Environ. Res. Risk Assess.* 29, 623–635. doi: 10.1007/s00477-014-1016-y

Moustakas, A., and Evans, M. R. (2017). A big-data spatial, temporal and network analysis of bovine tuberculosis between wildlife (badgers) and cattle. *Stoch. Environ. Res. Risk Assess.* 31, 315–328. doi: 10.1007/s00477-016-1311-x

Niphadkar, M., and Nagendra, H. (2016). Remote sensing of invasive plants: incorporating functional traits into the picture. *Int. J. Remote Sens.* 37, 3074–3085. doi: 10.1080/01431161.2016.1193795

Ordonez, A., and Olff, H. (2013). Do alien plant species profit more from high resource supply than natives? A trait-based analysis. *Global Ecol. Biogeogr.* 22, 648–658. doi: 10.1111/geb.12019

Panitsa, M., Tzanoudakis, D., Triantis, K. A., and Sfenthourakis, S. (2006). Patterns of species richness on very small islands: the plants of the Aegean archipelago. *J. Biogeogr.* 33, 1223–1234. doi: 10.1111/j.1365-2699.2006.01481.x

Papanastasis, V. P., and Kazaklis, A., (1998). "Land use changes and conflicts in the mediterranean-type ecosystems of western Crete," in *Landscape Disturbance and Biodiversity in Mediterranean-Type Ecosystems,* eds P. Rundel, G. Montenegro and F. Jaksic (Berlin; Heidelberg: Springer), 141–154.

Peltzer, D. A., Kurokawa, H., and Wardle, D. A. (2016). Soil fertility and disturbance interact to drive contrasting responses of co-occurring native and nonnative species. *Ecology* 97, 515–529. doi: 10.1890/15-0298.1

Planty-Tabacchi, A.-M., Tabacchi, E., Naiman, R. J., Deferrari, C. and Décamps, H. (1996). Invasibility of species-rich communities in riparian zones. *Conserv. Biol.* 10, 598–607. doi: 10.1046/j.1523-1739.1996.10020598.x

R Development Core Team (2016). *R: A Language and Environment for Statistical Computing*. Vienna: R Foundation for Statistical Computing. ISBN 3-900051-07-0.

Reaser, J. K., Meyerson, L. A., Cronk, Q., and De Poorter, M. (2007). Ecological and socioeconomic impacts of invasive alien species in island ecosystems. *Environ. Conserv.* 34, 98–111. doi: 10.1017/S0376892907003815

Reis, S., Seto, E., Northcross, A., Quinn, N. W., Convertino, M., Jones, R. L., et al. (2015). Integrating modelling and smart sensors for environmental and human health. *Environ. Model. Softw.* 74, 238–246. doi: 10.1016/j.envsoft.2015.06.003

Richardson, D. M., and Pyšek, P. (2006). Plant invasions: merging the concepts of species invasiveness and community invasibility. *Prog. Phys. Geogr.* 30, 409–431. doi: 10.1191/0309133306pp490pr

Robinson, G. R., Quinn, J. F., and Stanton, M. L. (1995). Invasibility of Experimental Habitat Islands in a California Winter Annual Grassland. *Ecology* 76, 786–794. doi: 10.2307/1939344

Rodriguez, L. F. (2006). Can invasive species facilitate native species? evidence of how, when, and why these impacts occur. *Biol. Invasions* 8, 927–939. doi: 10.1007/s10530-005-5103-3

Ross, L. C., Lambdon, P. W., and Hulme, P. E. (2008). Disentangling the roles of climate, propagule pressure and land use on the current and potential elevational distribution of the invasive weed Oxalis pes-caprae L. on Crete. *Perspect. Plant Ecol. Evol. Syst.* 10, 251–258. doi: 10.1016/j.ppees.2008.06.001

Sarris, A., Maniadakis, M., Lazaridou, O., and Pirintsos. S. A. (2005). Studying land use patterns in Crete Island, Greece, through a time sequence of landsat images and mapping vegetation patterns. *WSEAS Trans. Envir. Dev.* 1, 272–279. Available online at: http://emeric.ims.forth.gr/Landuse%20patterns-Vegetation%20mapping%20of%20Crete.pdf

Simberloff, D. (2009). The role of propagule pressure in biological invasions. *Annu. Rev. Ecol. Evol. Syst.* 40, 81–102. doi: 10.1146/annurev.ecolsys.110308.120304

Souza, L., Bunn, W. A., Simberloff, D., Lawton, R. M., and Sanders, N. J. (2011). Biotic and abiotic influences on native and exotic richness relationship across spatial scales: favourable environments for native species are highly invasible. *Funct. Ecol.* 25, 1106–1112. doi: 10.1111/j.1365-2435.2011.01857.x

Stohlgren, T. J., Barnett, D. T., and Kartesz, J. T. (2003). The rich get richer: patterns of plant invasions in the United States. *Front. Ecol. Environ.* 1, 11–14. doi: 10.1890/1540-9295(2003)001[0011:TRGRPO]2.0.CO;2

Stohlgren, T. J., Binkley, D., Chong, G. W., Kalkhan, M. A., Schell, L. D., Bull, K. A., et al. (1999). Exotic plant species invade hot spots of native plant diversity. *Ecol. Monogr.* 69, 25–46. doi: 10.1890/0012-9615(1999)069[0025:EPSIHS]2.0.CO;2

Stohlgren, T. J., Jarnevich, C., Chong, G. W., and Evangelista, P. H. (2006). Scale and plant invasions: a theory of biotic acceptance. *Preslia* 78, 405–426. Available online at: http://www.preslia.cz/P064CSto.pdf

Thomson, D. (2005). Measuring the effects of invasive species on the demography of a rare endemic plant. *Biol. Invasions* 7, 615–624. doi: 10.1007/s10530-004-5853-3

Trigas, P., Panitsa, M., and Tsiftsis, S. (2013). Elevational gradient of vascular plant species richness and endemism in crete–the effect of post-isolation mountain uplift on a continental island system. *PLoS ONE* 8:e59425. doi: 10.1371/journal.pone.0059425

Turland, N. J., Chilton, L., and Press, J. R. (1993). *Flora of the Cretan Area: Annotated Checklist and Atlas*. London: The Natural History Museum, HMSO.

Vetaas, O. R., and Grytnes, J.-A. (2002). Distribution of vascular plant species richness and endemic richness along the Himalayan elevation gradient in Nepal. *Global Ecol. Biogeogr.* 11, 291–301. doi: 10.1046/j.1466-822X.2002.00297.x

Vilà, M., Espinar, J. L., Hejda, M., Hulme, P. E., Jarošík, V., Maron, J. L., et al. (2011). Ecological impacts of invasive alien plants: a meta-analysis of their effects on species, communities and ecosystems. *Ecol. Lett.* 14, 702–708. doi: 10.1111/j.1461-0248.2011.01628.x

Vilà, M., and López-Darias, M. (2006). Contrasting biogeography of endemic and alien terrestrial species in the Canary Islands. *Orsis* 21, 91–101. Available online at: https://ddd.uab.cat/pub/orsis/02134039v21/02134039v21p91.pdf

Vogiatzakis, I. N., Griffiths, G. H., and Mannion, A. M. (2003). Environmental factors and vegetation composition, Lefka Ori massif, Crete, *S. Aegean. Global Ecol. Biogeogr.* 12, 131–146. doi: 10.1046/j.1466-822X.2003.00021.x

Wallace, J. M., and Prather, T. S. (2016). Invasive spread dynamics of *Anthriscus caucalis* at an ecosystem scale: propagule pressure, grazing disturbance and plant community susceptibility in canyon grasslands. *Biol. Invasions* 18, 145–157. doi: 10.1007/s10530-015-0997-x

Wood, S. (2006). G*eneralized Additive Models: An Introduction with R*. Boca Raton, FL: Chapman & Hall/CRC.

Zhang, Z., Xie, Y., and Wu, Y. (2006). Human disturbance, climate and biodiversity determine biological invasion at a regional scale. *Integr. Zool.* 1, 130–138. doi: 10.1111/j.1749-4877.2006.00029.x

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers
in Applied Mathematics and Statistics

ORIGINAL RESEARCH
published: 29 August 2017
doi: 10.3389/fams.2017.00017

Combining Phylogenetic and Occurrence Information for Risk Assessment of Pest and Pathogen Interactions with Host Plants

Ángel L. Robles-Fernández [1, 2] *and Andrés Lira-Noriega* [2, 3]*

[1] *Facultad de Física, Universidad Veracruzana, Xalapa, Mexico,* [2] *Instituto de Ecología, A.C., Red de Estudios Moleculares Avanzados, Xalapa, Mexico,* [3] *Consejo Nacional de Ciencia y Tecnología (CONACYT), Mexico City, Mexico*

Phytosanitary agencies conduct plant biosecurity activities, including early detection of potential introduction pathways, to improve control and eradication of pest and pathogen incursions. For such actions, analytical tools based on solid scientific knowledge regarding plant-pest or pathogen relationships for pest risk assessment are needed. Recent evidence indicating that closely related species share a higher chance of becoming infected or attacked by pests has allowed the identification of taxa with different degrees of vulnerability. Here, we use information readily available online about pest-host interactions and their geographic distributions, in combination with host phylogenetic reconstructions, to estimate a pest-host interaction (in some cases infection) index in geographic space as a more comprehensive, spatially explicit tool for risk assessment. We demonstrate this protocol using phylogenetic relationships for 20 beetle species and 235 host plant genera: first, we estimate the probability of a host sharing pests, and second, we project the index in geographic space. Overall, the predictions allow identification of the pest-host interaction type (e.g., generalist or specialist), which is largely determined by both host range and phylogenetic constraints. Furthermore, the results can be valuable in terms of identifying hotspots where pests and vulnerable hosts interact. This knowledge is useful for anticipating biological invasions or spreading of disease. We suggest that our understanding of biotic interactions will improve after combining information from multiple dimensions of biodiversity at multiple scales (e.g., phylogenetic signal and host-vector-pathogen geographic distribution).

Keywords: ambrosia beetle, biological invasions, biotic interaction, host range, logistic regression, phylogenetic distance, phytosanitary risk assessment

OPEN ACCESS

Edited by:
Stelios Katsanevakis,
University of the Aegean, Greece

Reviewed by:
Roger D. Magarey,
North Carolina State University,
United States
Helen E. Roy,
Centre for Ecology and Hydrology,
United Kingdom

Correspondence:
Andrés Lira-Noriega
aliranoriega@gmail.com

Specialty section:
This article was submitted to
Environmental Informatics,
a section of the journal
Frontiers in Applied Mathematics and Statistics

Received: *08 May 2017*
Accepted: *09 August 2017*
Published: *29 August 2017*

Citation:
Robles-Fernández ÁL and Lira-Noriega A (2017) Combining Phylogenetic and Occurrence Information for Risk Assessment of Pest and Pathogen Interactions with Host Plants.
Front. Appl. Math. Stat. 3:17.
doi: 10.3389/fams.2017.00017

1. INTRODUCTION

Understanding the geographic distribution and evolutionary ecology of plant pests and pathogens represents an advance in phytosanitary risk assessment. While pests and pathogens can arrive at new locations through natural dispersal, or be introduced for the purpose of biological control or accidentally, or even by hybridization *in situ* [1], their relationships with specific hosts highlights their potential to interact with different hosts, which is largely determined by their phylogenetic constraints. Typically, phytosanitary agencies provide plant biosecurity at regional or national levels through eradication or control of new pests via preventive activities such as quarantining, detection of illegal introduction through customs, and other local measures, once the problem is detected. However, although pest control and phytosanitary risk assessment require a thorough analysis of the areas of distribution of pests or pathogens, the environmental conditions in which they develop, and the hosts with which they are associated, resources available for carrying out such analyses are normally insufficient. A more effective alternative for an efficient phytosanitary response should be based on analytical tools that incorporate sound scientific knowledge regarding plant-pest relationships to help identify which pests pose risks [2, 3]. Two of the most informative sources of information (i.e., dimensions of biodiversity [4, 5]) for phytosanitary risk assessment include the geographic distribution of the pests and their associated hosts and phylogenetic distance among hosts as a surrogate of vulnerability, given that phylogenetically close species or genera share a higher probability of being affected by the same pest [6]. It is possible to encode information about phylogenetic distances among species in such a way that it can be shown geographically [7], thus providing a spatially explicit risk assessment of the plant-pest interactions. On this theoretical basis, it is possible to calculate the probability of infection by a pathogen or attack by a pest, either through its geographical distribution or by using statistical methods to estimate the incidence of pests on a series of hosts of different phylogenetic distances. The combination of these two sources of information regarding pest/pathogen-host interactions would provide a more comprehensive and integrated biogeographic approach to risk analysis.

In general, information currently available regarding the interaction of plants with pests or pathogens, or even insects, is scarce and mainly pertains to a few sites. However, it is possible to use information about the number of known hosts and the phylogenetic distance between known hosts and other species of interest to make inferences regarding plant-taxon interactions at different taxonomic or phylogenetic levels [6]. Gilbert et al. [6] used the Global Pest and Disease Database of the Plant Protection and Quarantine Division of the Animal and Plant Health Inspection Service of the United States Department of Agriculture (APHIS-PPQ) database to extract all of the recorded plant pests from 210 genera of flowering plants, and evaluate the strength of the phylogenetic signal in terms of host range for nine major groups of plant pests and pathogens. They found clear statistical patterns in terms of the likelihood of different plant genera sharing a common pest. On the other hand, primary biodiversity data of species distributions for major taxa can be readily accessed via several online databases, such as the Global Biodiversity Information Facility (GBIF)[1], and can be used simply as occurrence information with which to conduct spatial queries or as input data for obtaining potential distributional ranges from species distribution modeling. As distributional data from such databases present spatial and taxonomic biases and gaps, they must therefore be used with some caution [8].

Here, we propose an approach by which to: (1) assess the probability that a source host shares a pest with a target host, considering the phylogenetic distance between them, and (2) use this model to project the intensity of the expected likelihood in geographic space given the distribution of host taxa in what could be interpreted as a spatial index of interaction or risk to host plants. This procedure represents an efficient method for performing a geographical assessment, especially when interpretation depends on the interactions of the pest or pathogen with host plants or any other agent (e.g., insects) found to be associated with the host plants. We apply this protocol to a set of ambrosia beetles (Scolytinae: Xyleborina) and their associated hosts to illustrate the efficacy of this assessment and evaluate its similarity to empirical evidence collected previously [6, 9].

2. METHODS

2.1. Beetle and Plant Host Incidence Matrix

We derived an incidence matrix I based on information pertaining to 23 ambrosia beetle species and their associated host plant species by retrieving information from one of the most comprehensive databases on bark and ambrosia beetles, curated by Thomas Atkinson[2]. These 23 species were selected because of their phylogenetic relationship to *Xyleborus glabratus*, an invasive alien species in North America and vector of the fungus *Raffaelea lauricola*, which causes Laurel Wilt Disease and is a major agent for tree mortality for species of Lauraceae [10]. In general, these data coincided with recently published information on the phylogenetic relationships of ambrosia beetles [11]. The taxonomy of beetle species to *X. glabratus* was checked using information validated for GBIF and NCBI with the R package taxize [12].

Using information from the same database, we found 372 host plant species associated with this set of beetle species. From this group of host plant species, we summarized information at the genus level (243 genera), of which we kept only those genera with phylogenetic information valid for the Angiosperm Phylogeny Group (APG III; [13] for reconstructing a phylogenetic hypothesis (235 genera in total; see below). A final incidence matrix $\mathrm{I} = a_{ij}$ was then obtained with 20 beetle species for rows i and 235 host plant genera for columns j, coding it as 1 if the beetle-plant association was reported as present and 0 otherwise. We considered only beetle species with three or more plant associations, since a lower number of incidences could lead to unreliable results in the linear regression ([6]; see below).

$$\mathrm{I} = \left\{ a_{ij} | a_{ij} = \begin{cases} 1 & \text{if host is infected} \\ 0 & \text{if} \quad \text{else} \end{cases} \right\} \tag{1}$$

2.2. Plant Host Phylogenetic Distance Matrix

For the 235 host plant genera, an ultrametric phylogenetic tree was obtained from Phylomatic through the R package brranching

---

[1] http://www.gbif.org

[2] http://www.barkbeetles.info/

[14]. This tree was then imported into Phylomatic, where branch lengths were estimated using the bladj function [15]. This step uses information from the APG III supertree [13] and the dated nodes given by Wikstrom [16]. From this phylogenetic tree, we then obtained a symmetric real-valued dissimilarity matrix $D_1 = [d_{ij}]$ of order $n = 235$, in which $d_{ii} = 0$ for all $1 \leq i \leq n$ and $d_{ij} \geq 0$ for all $1 \leq i < j \leq n$ for each n plant genus with the R package ape [17] and then transformed the matrix as $D = log_{10}(D_1 + 1)$, following Gilbert et al. [6].

2.3. Beetle-Host Association Probability Matrix

To obtain the probability of a host plant genus being affected by, or found in association with, a particular beetle species (or all of the beetle species), we calculated a logistic regression relating the host-beetle incidences in matrix I to the host phylogenetic distances in matrix D, using phylogenetic distance as the independent variable. To this end, we first constructed a two-column array between I and D as follows: for each row in I, we randomly selected one interacting host (known as the source host), and then selected the remaining host genera (target hosts) iteratively at random. We repeated this process for each row in I, adding the information corresponding to each new row to the two-column array until we had completed the entire set of beetle species. A matrix of probabilities P was then obtained by applying the regression coefficients to the logistic transformation of I (see Section 2.2). In Equation (2), β_0 and β_1 were obtained by repeating the previous procedure 1,000 times and selecting both the mean intercepts (β_0) and slope coefficients (β_1) of these regressions.

$$P = \frac{1}{1 + e^{-\beta_0 + \beta_1 D}} \qquad (2)$$

2.4. Presence-Absence Matrices Of Host Plants

A presence-absence matrix summarizes information pertaining to the geographic distribution of every host plant species over a regular grid for any given region, with presence coded as 1 and absence as 0. We defined $M = [m_{ij}]$ in which $m_{ij} = 1$ if the host genera was present and $m_{ij} = 0$ where absent. If this occurred, we considered two cases with which to show the host plant genera distribution on a regular grid of 1 degree spatial resolution: one matrix containing information on host plant genera, derived from the list of host plant species in the database of Atkinson (called M_1), and a second matrix corresponding to host plant genera in the phylogenetic tree but showing occurrences for each genus directly from GBIF using the R package spocc [18] (called M_2). For M_1, all 372 host plant species were used to obtain occurrence data from GBIF, after which we grouped information at the genus level and merged it with the genera in the phylogenetic distance matrix D. This procedure rendered a matrix of $m = 232$ host plant genera by $n = 62,640$ sites. For M_2, the occurrence data from GBIF rendered a matrix of $m = 203$ host plant genera by $n = 62,640$ sites. We used both matrices M_1 and M_2 to illustrate the fact that the occurrence data could also be biased or incomplete when using different avenues by which to retrieve it. This problem is potentially common when handling resources available on different databases or using different search engines (e.g., accessing GBIF directly online or through the R package spocc).

2.5. Interaction or Infection Spatial Index and Risk Maps

We obtained an interaction or infection matrix $G = PM$ by multiplying matrix P by matrix M_1 or M_2. Matrix G contains information about the probability that a given source host can share a beetle with other target hosts when they are present at a given site. We suppose that each interaction or infection event could occur regardless of which host is the source, an ideal scenario that simplifies the process of the interaction or infection. The interaction or infection index per site $g = \sum_{j=1}^{m} [g_{ij}]$ was then obtained by summing each column in G. Finally, to spatialize this index vector, we assigned the longitude and latitude of every cell in the 1 degree grid to build a raster that portrays the index value over geographic space. This index was normalized and rescaled with cumulative and generalized logistic transformation using the R package bossMaps [19], allowing us to compare among different outputs (e.g., different sets of species or genera); this index is equivalent to a suitability index. We also converted the logistic output to a binary map, using a threshold of 0.5 (see Supplementary Material).

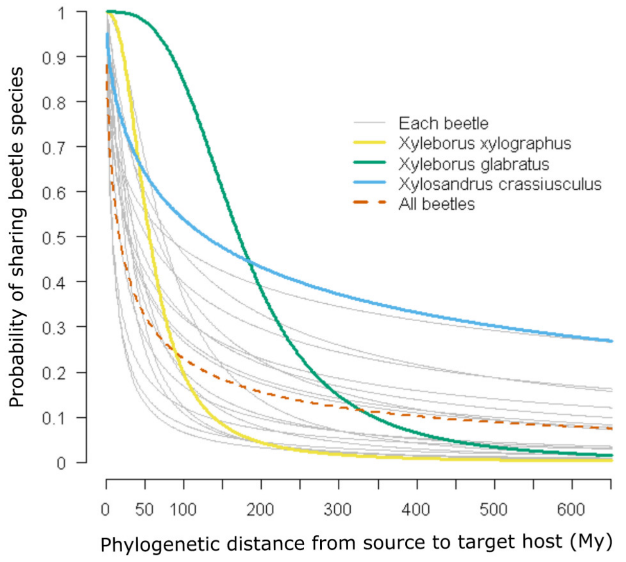


FIGURE 1 | Phylogenetic signal and probability of host plant genera sharing beetle species. Curves are predicted from logistic regressions using the coefficients given in **Table 1**. Four relationships are highlighted: a narrow host range and phylogenetically constrained species (*Xyleborus xylographus*), a wide host range and phylogenetically constrained species (*Xyleborus glabratus*), a wide host range and phylogenetically dispersed species (*Xylosandrus crassiusculus*), and all beetle species together.

TABLE 1 | Logistic regression coefficients and confidence intervals.

Beetle species	β_0	2.5%	97.5%	β_1	2.5%	97.5%
Ambrosiodmus obliquus	2.005	−3.5225	7.5325	−2.1993	−4.3305	−0.0682
Ambrosiodmus rubricollis	5.1398	−4.1651	14.4448	−3.3163	−6.8702	0.2377
Ambrosiodmus tachygraphus	4.6856	−9.1553	18.5264	−3.3991	−8.6864	1.8882
Xyleborus obesus	2.3697	−4.2957	9.0352	−2.5059	−5.0497	0.038
Cnestus mutilatus	3.4716	−2.5458	9.489	−2.4619	−4.7168	−0.207
Coptoborus pseudotenuis	7.7836	−3.7667	19.3338	−4.192	−8.639	0.2551
Euwallacea fornicatus	3.014	−3.2981	9.3261	−1.8585	−4.2341	0.5171
Euwallacea validus	4.7687	−1.7713	11.3087	−2.6016	−5.0338	−0.1695
Sampsonius dampfi	5.8214	−6.8643	18.5071	−3.8804	−8.7952	1.0343
Xyleborinus saxeseni	2.6893	−3.0522	8.4309	−1.8111	−3.9597	0.3375
Xyleborus affinis	2.1088	−2.9366	7.1543	−1.1084	−2.9856	0.7688
Xyleborus ferrugineus	2.6372	−2.3476	7.622	−1.5189	−3.395	0.3571
Xyleborus glabratus	15.941	7.6353	24.2468	−7.1511	−10.3155	−3.9866
Xyleborus posticus	1.7296	−3.4087	6.8678	−1.8582	−3.8101	0.0936
Xyleborus seriatus	2.6659	−3.6384	8.9702	−2.1272	−4.4742	0.2197
Xyleborus xylographus	9.6907	−4.2267	23.6081	−5.5552	−11.0613	−0.0491
Xylosandrus compactus	3.3348	−3.6878	10.3575	−2.0683	−4.6553	0.5188
Xylosandrus crassiusculus	2.9637	−2.7676	8.695	−1.4082	−3.5437	0.7273
Xyleborus germanus	2.319	−2.9065	7.5446	−1.5299	−3.4579	0.3982
Xylosandrus morigerus	3.5355	−2.0179	9.0889	−1.8522	−3.9433	0.2389

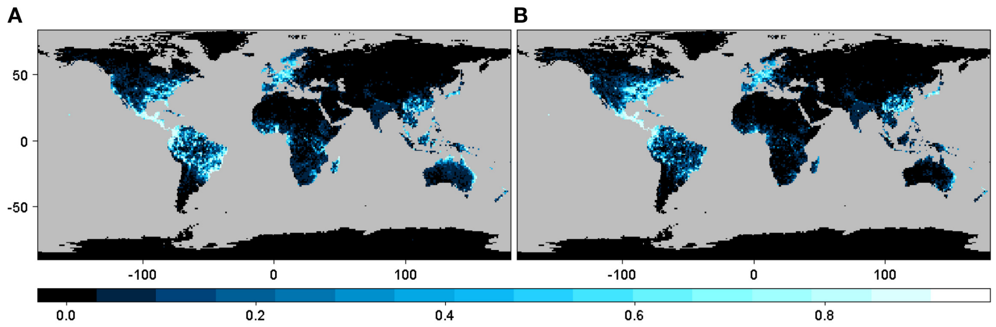


FIGURE 2 | Spatial projection of the interaction or infection index summarized at 1-degree spatial resolution for the case with all of the beetles pooled together using GBIF occurrence information from **(A)** host plant genera derived from the list of host plant species in the bark beetles database (M_1) and **(B)** host plant genera in the phylogenetic tree after obtaining the occurrences for each genus directly using the R package spocc (M_2).

2.6. Relationship between Host Range, Host Richness and Interaction or Infection Spatial Index

To explore the relationship between host range (i.e., the number of host genera interacting with a beetle species) and the correlation between host richness (i.e., number of host genera per cell in the grid) and the interaction or infection index, we obtained the correlation coefficients r_n between the interaction or infection index for each beetle species g_n ($n = 20$ cases) and host richness S ($r_{g_n,S} = corr(g_n, S)$). These correlation coefficients were then used to perform a linear regression with a negative exponential transformation of the host range (Equation 3). Host richness was estimated by $S = \sum_{m=1}^{j} a_{ij}$ for m_{ij} both for M_1 and M_2, respectively. Host range was obtained from $H = \sum_{m=1}^{j} a_{ij}$ for a_{ij} elements of **I**. All methods were implemented in the R package geotax available through GitHub[3].

$$r = \alpha_0 + \alpha_1 e^{-H} \tag{3}$$

3. RESULTS

Figure 1 shows the probability of a host plant genus sharing a beetle interaction determined by the phylogenetic distance from the source host using the coefficients of the logistic regression (**Table 1**). Particular cases are highlighted in the figure, in which the red dotted line corresponds to the case where all beetles

---

[3] https://github.com/alrobles/geotax

are included in the analysis; three other cases were selected to show relationships in which the beetle corresponds to (1) a narrow host range and phylogenetically constrained species (*Xyleborus xylographus*), (2) an intermediate host range and phylogenetically constrained species (*Xyleborus glabratus*), and (3) a wide host range and phylogenetically dispersed species (*Xylosandrus crassiusculus*).

The geographic expression of the interaction or infection index in the case of all beetles pooled together derived from both M_1 and M_2 shows approximately the same pattern, although some differences in the pattern distribution are apparent depending on the particular host plant distribution matrix used (**Figure 2**). These maps indicate that the highest probability of interaction or infection occurs over large areas worldwide, with the highest intensity of this process occurring in parts of Neotropical America and the eastern and western coasts of North America, Central Europe, smaller portions of sub-Saharan Africa and Madagascar, Southeast Asia, and eastern and northern portions of Australia. These patterns corresponded in large part to the distribution of the 235 host genera worldwide. When the distribution of the index was mapped over the binary map, we observed a reduction in the distribution of the interaction or infection index (Figure S1).

The geographic distribution of the interaction or infection index for the three particular cases of phylogenetic constraint and host range depicted different regions depending on the beetle species and host plant genera distribution matrix, although these differences were not so marked between M_1 and M_2 (**Figure 3**). In the narrow host range and phylogenetically constrained species (*Xyleborus xylographus*), the intensity of the interaction was highest in eastern North America, Central Europe, and parts of Southeast Asia. For the intermediate host range and phylogenetically constrained species (*Xyleborus glabratus*), highest intensity was in eastern North America, from the northern limit of the Neotropics through Central America, and in northern and eastern South America, as well as in eastern China, Southeast Asia, and eastern Australia. In the wide host range and phylogenetically dispersed species (*Xylosandrus crassiusculus*), the distribution of the interaction index intensity was similar to *X. glabratus*.

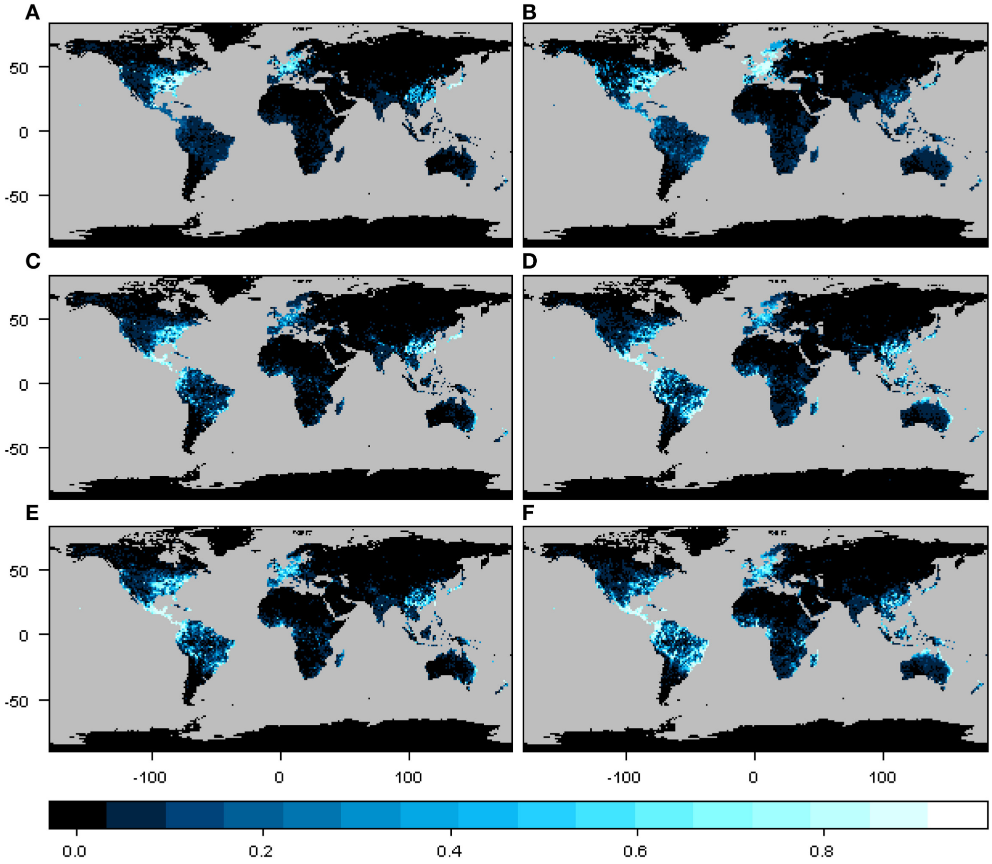


FIGURE 3 | Spatial interaction or infection index for three cases of host range size and different levels of phylogenetic constraint. Maps in **(A,C,E)** correspond to the interaction or infection index for host plant genera from the original host-beetle database (M_1), and maps in **(B,D,F)** correspond to the interaction or infection index for host plant genera from the phylogenetic tree (M_2). The three cases correspond to a narrow host range and phylogenetically constrained species (*Xyleborus xylographus*), a wide host range vector and phylogenetically constrained species (*Xyleborus glabratus*) and a wide host range and phylogenetically dispersed species (*Xylosandrus crassiusculus*). Maps summarize information at 1-degree spatial resolution.

We observed a linear relationship (Equation 3) between the transformed host range size H and the correlation coefficient r between the interaction or infection index and host richness. The values to calculate the relationship are shown in **Table 2**. This pattern was constant when the correlation was conducted for both host plant genera matrices M_1 and M_2 (**Figure 4**). This result showed that, with a small number of interacting hosts, the model had a lower correlation with host richness; however, when the host range was large, the correlation with host richness grew asymptotically to one. It is thus possible to characterize quantitatively the nature of the interaction (according to host range and phylogenetic constraint) in relation to a set of pests or interacting agents and their hosts.

4. DISCUSSION

Prediction of biological interactions and invasions is a challenge. From a phytosanitary perspective, meeting this challenge would yield the potential to anticipate the ecological and evolutionary dynamics of species and the potential for alteration of native or cultivated ecosystems. Here, we show that combining information from the ecological and phylogenetic relationships of pest species and geographic distributions of known and potential hosts offers an opportunity to map risk of problems even before a new pest is recognized. This process could also be used to anticipate the potential for species invasions [20] or outbreaks of disease related to novel pathogens [21].

Development of interaction or infection predictions involves two main steps: The first is to estimate the likelihood of sharing a pest from simple linear regression models *sensu* [6]. The second is to use that likelihood to predict the geographic patterns of interactions, considering the distribution of known hosts of the species, or those that could potentially interact with the species, given their phylogenetic proximity to the interacting agents (i.e., vectors, pests, etc.). The first step of this process allows estimation of the probability of sharing a host, considering different levels of phylogenetic proximity among hosts and host range size. This result is therefore very informative from an ecological and evolutionary perspective in terms of evaluation of patterns of infection [9]. The second step projects the likelihood of interaction over geographic space, presenting spatially explicit predictions of the interaction between the agents and their hosts.

While we know of no previous attempt to represent this combination of factors geographically, some previous efforts to combine ecological, phylogenetic, and geographic information have been made. For example, Leibold et al. [7] coded phylogenetic nodes per site to obtain geographical patterns of the phylogenetic signal, but without considering the interaction process. Furthermore, in a recent study Morales-Castilla et al. [22] summarized different frameworks to infer biotic interactions from different biodiversity dimensions, one of which included a combination of phylogenetic information

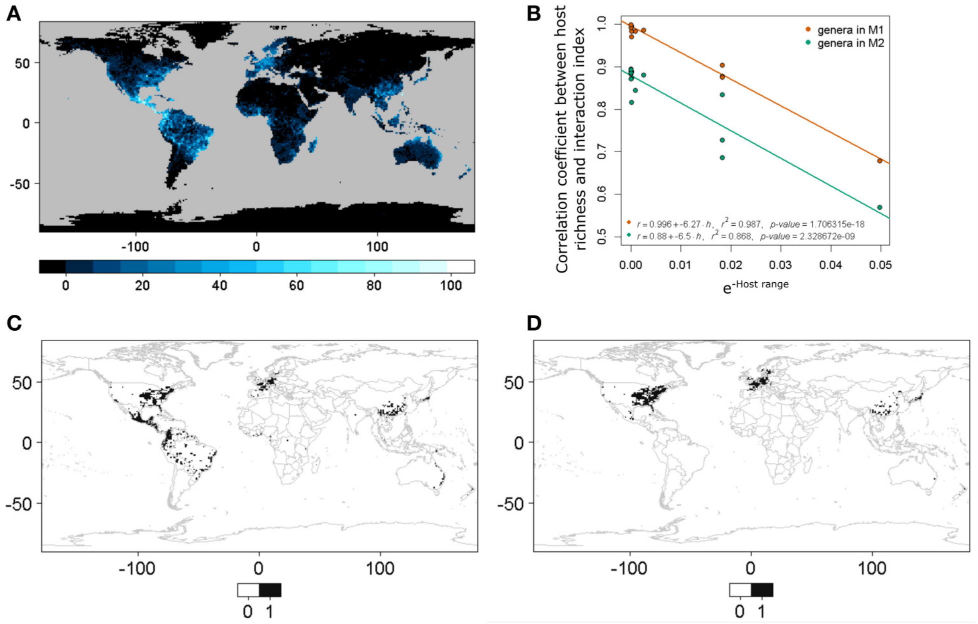


FIGURE 4 | (A) World host richness. This corresponds to host plant genera in M_1. **(B)** Linear model between host range (negative exponential transformation, also see Equation 3; a small value implies a larger host range and vice versa) and the correlation coefficients between host richness and the interaction or infection spatial index. M_1 corresponds to host plant genera derived from the list of host plant species in the bark beetles database and M_2 to host plant genera in the phylogenetic tree after obtaining the occurrences for each genera directly using the R package spocc. **(C)** Binary predictive model for *Xylosandrus crassiusculus*, a generalist bark beetle species. **(D)** Binary predictive model for *Xyleborus xylographus*, a specialist bark beetle species. Note that cases **(C,D)** correspond to extremes in **(B)**, and that **(C)** has a higher correlation with species host richness than **(D)**.

and species distributions. However, none of these frameworks transferred or projected the inferred interactions to geographic space. We suggest that this step is critical in anticipating potential interactions; from our analysis, we can conclude that the potential geographic distribution depends on, or is a consequence of, the number and type of interactions (i.e., according to host range and level of phylogenetic constraint; also see [23, 24]). An extension of this approach could be to improve single species distribution models [25] by including both the customary environmental information and species interactions. Moreover, since this index may be interpreted as a suitability index for predicting ecological interactions, it could also be interpreted as summarizing the parts of the realized ecological niche of the species related to the Eltonian ecological niche [26, 27]. Given the fact that biotic interactions are difficult to predict and quantify at geographic scales [28], this information is crucial for predicting the geographic distributions of species [29, 30].

In our example using ambrosia beetles, we detected at least three main cases of potential interaction between the likelihood of sharing an interaction and phylogenetic distance among hosts. These situations were the narrow host range and phylogenetically constrained species (*Xyleborus xylographus*), narrow to intermediate host range and phylogenetically constrained species (*Xyleborus glabratus*) and wide host range and phylogenetically dispersed species (*Xylosandrus crassiusculus*). These cases show a different pattern in terms of the likelihood of sharing an interaction, which is quantified from the coefficients of the linear regression (see **Figure 1** and **Table 1**). Moreover, these individual patterns correspond well with the host range known for these beetle species from the original database. Prediction of these patterns had been suggested by Gilbert et al. [6] and other authors have found similar trends [31]; however, those authors did not show the geographic pattern of the relationship with actual or potential hosts.

TABLE 2 | Host range for each beetle species and correlation coefficients between host plant genera richness and the interaction or infection index from M_1 and M_2.

	Host range	r_{M_1}	r_{M_2}
Ambrosiodmus obliquus	6	0.8814	0.9859
Ambrosiodmus rubricollis	7	0.8451	0.9841
Ambrosiodmus tachygraphus	4	0.7278	0.8787
Xyleborus obesus	4	0.6861	0.8758
Cnestus mutillatus	10	0.873	0.997
Coptoborus pseudotenuis	10	0.8897	0.9843
Euwallacea fornicatus	31	0.8952	0.9988
Euwallacea validus	22	0.8836	0.9988
Sampsonius dampfi	4	0.8339	0.9042
Xyleborinus saxeseni	25	0.8876	0.999
Xyleborus affinis	70	0.8918	0.9999
Xyleborus ferrugineus	47	0.8927	0.9995
Xyleborus glabratus	9	0.8163	0.9713
Xyleborus posticus	10	0.8873	0.9945
Xyleborus seriatus	11	0.8714	0.993
Xyleborus xylographus	3	0.5687	0.6785
Xylosandrus compactus	22	0.8882	0.9996
Xylosandrus crassiusculus	73	0.8917	0.9999
Xyleborus germanus	32	0.8859	0.9996
Xylosandrus morigerus	48	0.8939	0.9992

We conclude from these single-species relationships that, to obtain a statistically significant linear regression model or a relationship with less uncertainty, it is necessary to have several observed interactions with different hosts or use quantitative impact data [32]. In our single-species examples, we excluded cases in which the beetles had two or fewer host plant species. In addition, the relationship pattern observed with all beetles pooled together was frequently different from that of single-species cases, thus suggesting that useful information exists at the single-species level that differs from that of the whole set of species. Interestingly, when we project the interaction or infection index for the single-species cases, it is clear that potential interaction with hosts depends largely on their geographic distribution and also that the differences in the predictions for the individual cases are not particularly marked. This outcome could be due to the fact that these beetle species are phylogenetically related and, in some cases, share host genera that have similar geographic distributions. However, we would expect this result to change for different taxonomic groups or sets of species.

Overall, the geographic distribution patterns of the interaction or infection index coincide with the distributions of host genera in both the native and invaded areas of the beetles. The most recently invaded areas for some of these species in North America (i.e., *Xyleborus glabratus* and *Euwallacea* sp.) coincide with a pattern of invasion with high intensity values for the interaction or infection index. Areas that are not yet invaded but where the plant host genera are present show high invasive potential, as predicted by other studies. For example, the vector of the fungus *Raffaelea lauricola*, *Xyleborus glabratus* which causes the Laurel Wilt Disease, has shown relatively fast expansion in eastern North America following its introduction from Asia in 2002 [33]. The interaction or infection index for this species predicted not only its native range in Asia, but also its currently invaded range in the United States. This species has advanced from Georgia to Texas in a period of 15 years, affecting species from the Lauraceae family and impacting native and cultivated plantations[4]. However, its invasive potential expands southwards to regions in Mexico, Central, and South America, where Lauraceae species are diverse and abundant, and where cultivated plantations of avocado (*Persea americana*) are also an economically important crop. Hence, the interaction or infection index appears to predict what is known about the current distribution for these species and can be a valuable tool for anticipating host ranges and areas with potential for invasion.

The fact that we observed a linear relationship between the host range size and the correlation between the interaction or infection index and host richness suggests that host range size and host diversities are important drivers of ecological and evolutionary interactions. Overall, there is a higher intensity of the interaction or infection index between beetles and hosts

---

[4] https://www.fs.usda.gov/Internet/FSE_DOCUMENTS/fseprd523011.pdf

in more host-diverse regions. This pattern can be useful for interpretation and regionalization of the expected distribution of interactions, considering the diversity of hosts. However, an important difference exists in the results depending on the amount of information used (e.g., whether plant host genera are derived from M_1 or M_2). We suggest that host richness alone does not provide sufficiently good prediction and that converting the index to a binary map allows for improved delimitation of the areas at higher risk (e.g., **Figure 4**). This interaction or infection index projected in geographic space provides a manner by which to re-scale the process of infection using both host richness and phylogenetic distance. It is also important to consider that while there is currently more primary biodiversity information (i.e., information on species occurrences) available from which to address new questions from the perspective of biodiversity informatics, variation in both the availability and quality of data remains an important problem to be resolved [34].

In summary, the combination of different biodiversity dimensions is a necessary avenue by which to understand patterns and processes in biodiversity [4, 34–36]. While further theoretical background is still necessary to understand the extent and limitations of combining phylogenetic and distributional information for anticipating biotic interactions, empirical evidence indicates that this tool is promising for confirming actual and vulnerable hosts. Implementation of this relationship in easy-to-use algorithms, such as we present here, represents a step forward toward evaluating risk in phytosanitary and biological invasion assessment, and in particular to project potential host-pest interactions over geographic space. This approach, however, would benefit from further fieldwork to test these models empirically [37]. We also expect that these implementations will be important for more quantitative approaches toward understanding ecological and evolutionary patterns and processes.

AUTHOR CONTRIBUTIONS

ÁR-F obtained the data, developed the algorithms and implemented the analysis. AL-N conceived the idea and design of the analysis and both authors reviewed the data and outcomes, and wrote the manuscript.

FUNDING

ÁR-F received partial support from the Consejo Veracruzano de Investigación Científica y Desarrollo Tecnológico (COVEICYDET) 5th Young Talent 2016 award for the analysis presented here. This project was supported by Fondo Sectorial para la Investigación, el Desarrollo y la Innovación Tecnológica Forestal (2015-C01-265677).

ACKNOWLEDGMENTS

We thank Enrique Ibarra-Laclette and Araceli Lamelas for making important suggestions on the assembly and quality of the phylogenies used in the analysis. Comments and suggestions from Greg Gilbert and Jorge Soberón were important for the development of this project. Reviews from the two assigned reviewers and from A. Townsend Peterson greatly improved the quality of the manuscript. Special thanks to Keith MacMillan for careful English revision.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: http://journal.frontiersin.org/article/10.3389/fams.2017.00017/full#supplementary-material

REFERENCES

1. Brasier CM. Rapid evolution of introduced plant pathogens via interspecific hybridization hybridization is leading to rapid evolution of Dutch elm disease and other fungal plant pathogens. *Bioscience* (2001) **51**:123–33. doi: 10.1641/0006-3568(2001)051[0123:REOIPP]2.0.CO;2
2. Campbell FT. The science of risk assessment for phytosanitary regulation and the impact of changing trade regulations: the approach to phytosanitary safeguards mandated by the World Trade Organization may hinder adoption of the most efficient methods to protect ecosystems from introductions of invasive species. *Bioscience* (2001) **51**:148–53. doi: 10.1641/0006-3568(2001)051[0148:TSORAF]2.0.CO;2
3. Magarey RD, Colunga-Garcia M, Fieselmann DA. Plant biosecurity in the United states: roles, responsibilities, and information needs. *Bioscience* (2009) **59**:875. doi: 10.1525/bio.2009.59.10.9
4. Stevens RD, Gavilanez, MM. Dimensionality of community structure: phylogenetic, morphological and functional perspectives along biodiversity and environmental gradients. *Ecography* (2015). **38**:861–75. doi: 10.1111/ecog.00847
5. Maclaurin J, Sterelny K. *What is Biodiversity?* Chicago, IL: University of Chicago Press (2008). doi: 10.7208/chicago/9780226500829.001.0001
6. Gilbert GS, Magarey R, Suiter K, Webb CO. Evolutionary tools for phytosanitary risk analysis: phylogenetic signal as a predictor of host range of plant pests and pathogens. *Evol Appl*. (2012) **5**:869–78. doi: 10.1111/j.1752-4571.2012.00265.x
7. Leibold MA, Economo EP, Peres-Neto P. Metacommunity phylogenetics: separating the roles of environmental filters and historical biogeography. *Ecol Lett*. (2010) **13**:1290–9. doi: 10.1111/j.1461-0248.2010.01523.x
8. Beck J, Böller M, Erhardt A, Schwanghart W. Spatial bias in the {GBIF} database and its effect on modeling species' geographic distributions. *Ecol Inform*. (2014) **19**:10–5. doi: 10.1016/j.ecoinf.2013.11.002
9. Parker IM, Saunders M, Bontrager M, Weitz AP, Hendricks R, Magarey R, et al. Phylogenetic structure and host abundance drive disease pressure in communities. *Nature* (2015) **520**:542–4. doi: 10.1038/nature14372
10. Wear, David N, Greis, John G (eds.). *The Southern Forest Futures Project: Technical Report*. General Technical Report SRS-GTR-178. Asheville, NC: USDA-Forest Service, Southern Research Station (2013). 542 p.
11. Gohli J, Kirkendall LR, Smith SM, Cognato AI, Hulcr J, Jordal BH. Biological factors contributing to bark and ambrosia beetle species diversification. *Evolution* (2017) **71**:1258–72. doi: 10.1111/evo.13219
12. Chamberlain S, Szocs E. taxize - taxonomic search and retrieval in R. *F1000Research* (2013). Available online at: http://f1000research.com/articles/2-191/v2
13. GROUP TAP. An update of the Angiosperm Phylogeny Group classification for the orders and families of flowering plants: APG III. *Botan J Linn Soc*. (2009) **161**:105–21. doi: 10.1111/j.1095-8339.2009.00996.x

14. Chamberlain S. *brranching: Fetch 'Phylogenies' from Many Sources*. (2016). R package version 0.2.0. Available online at: https://CRAN.R-project.org/package=brranching
15. Webb CO, Ackerly DD, Kembel SW. Phylocom: software for the analysis of phylogenetic community structure and trait evolution. *Bioinformatics* (2008) **24**:2098. doi: 10.1093/bioinformatics/btn358
16. Wikström N, Savolainen V, Chase MW. Evolution of the angiosperms: calibrating the family tree. *Proc R Soc Lond B Biol Sci*. (2001) **268**:2211–20. doi: 10.1098/rspb.2001.1782
17. Paradis E, Claude J, Strimmer K. APE: analyses of phylogenetics and evolution in R language. Bioinformatics (2004) **20**:289–90. doi: 10.1093/bioinformatics/btg412
18. Chamberlain S. *spocc: Interface to Species Occurrence Data Sources*. (2016). R package version 0.5.4. Available online at: https://CRAN.R-project.org/package=spocc
19. Merow C, Wilson AM, Jetz W. Integrating occurrence data and expert maps for improved species range predictions. *Glob Ecol Biogeogr*. (2017) **26**:243–58. doi: 10.1111/geb.12539
20. Worner S, Gevrey M. Modelling global insect pest species assemblages to determine risk of invasion. *J Appl Ecol*. (2006) **43**:858–67. doi: 10.1111/j.1365-2664.2006.01202.x
21. Poon L, Chu D, Chan K, Wong O, Ellis T, Leung Y, et al. Identification of a novel coronavirus in bats. *J Virol*. (2005) **79**:2001–9. doi: 10.1128/JVI.79.4.2001-2009.2005
22. Morales-Castilla I, Matias MG, Gravel D, Araújo MB. Inferring biotic interactions from proxies. *Trends Ecol Evol*. (2015) **30**:347–56. doi: 10.1016/j.tree.2015.03.014
23. Villalobos F, Rangel TF, Diniz-Filho JAF. Phylogenetic fields of species: cross-species patterns of phylogenetic structure and geographical coexistence. *Proc R Soc B* (2013) **280**:20122570. doi: 10.1098/rspb.2012.2570
24. Huang S, Bininda-Emonds OR, Stephens PR, Gittleman JL, Altizer S. Phylogenetically related and ecologically similar carnivores harbour similar parasite assemblages. *J Anim Ecol*. (2014) **83**:671–80. doi: 10.1111/1365-2656.12160
25. Morales-Castilla I, Davies TJ, Pearse WD, Peres-Neto P. Combining phylogeny and co-occurrence to improve single species distribution models. *Glob Ecol Biogeogr*. (2017) **26**:740–52. doi: 10.1111/geb.12580
26. Hutchinson GE. Cold spring harbor symposium on quantitative biology. *Concluding Remarks* (1957) **22**:415–27.
27. Soberón J. Grinnellian and Eltonian niches and geographic distributions of species. *Ecol Lett*. (2007) **10**:1115–23. doi: 10.1111/j.1461-0248.2007.01107.x
28. Peterson AT, Soberón J, Pearson R, Anderson R, Martínez-Meyer E, Nakamura M, et al. *Ecological Niches and Geographic Distributions (MPB-49)*. Princeton, NJ: Princeton University Press (2011).
29. Anderson RP. When and how should biotic interactions be considered in models of species niches and distributions? *J Biogeogr*. (2017) **44**:8–17. doi: 10.1111/jbi.12825
30. Wisz MS, Pottier J, Kissling WD, Pellissier L, Lenoir J, Damgaard CF, et al. The role of biotic interactions in shaping distributions and realised assemblages of species: implications for species distribution modelling. *Biol Rev*. (2013) **88**:15–30. doi: 10.1111/j.1469-185X.2012.00235.x
31. Cooper N, Griffin R, Franz M, Omotayo M, Nunn CL. Phylogenetic host specificity and understanding parasite sharing in primates. *Ecol Lett*. (2012) **15**:1370–7. doi: 10.1111/j.1461-0248.2012.01858.x
32. Gilbert GS, Briggs HM, Magarey R. The impact of plant enemies shows a phylogenetic signal. *PLoS ONE* (2015) **10**:e0123758. doi: 10.1371/journal.pone.0123758
33. Fraedrich S, Harrington T, Rabaglia R, Ulyshen M, Mayfield Iii A, Hanula J, et al. A fungal symbiont of the redbay ambrosia beetle causes a lethal wilt in redbay and other Lauraceae in the southeastern United States. *Plant Dis*. (2008) **92**:215–24. doi: 10.1094/PDIS-92-2-0215
34. Soberón J, Peterson T. Biodiversity informatics: managing and applying primary biodiversity data. *Philos Trans R Soc Lond B Biol Sci*. (2004) **359**:689–98. doi: 10.1098/rstb.2003.1439
35. Krishtalka L, Humphrey PS. Can natural history museums capture the future? *Bioscience* (2000) **50**:611–7. doi: 10.1641/0006-3568(2000)050[0611:CNHMCT]2.0.CO;2
36. Stephens PR, Altizer S, Smith KF, Alonso Aguirre A, Brown JH, Budischak SA, et al. The macroecology of infectious diseases: a new perspective on global-scale drivers of pathogen distributions and impacts. *Ecol Lett*. (2016) **19**:1159–71. doi: 10.1111/ele.12644
37. Schweizer D, Gilbert GS, Holl KD. Phylogenetic ecology applied to enrichment planting of tropical native tree species. *Forest Ecol Manage*. (2013) **297**:57–66. doi: 10.1016/j.foreco.2013.02.012

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers
in Ecology and Evolution

ORIGINAL RESEARCH
published: 29 November 2017
doi: 10.3389/fevo.2017.00149

Early Engagement of Stakeholders with Individual-Based Modeling Can Inform Research for Improving Invasive Species Management: The Round Goby as a Case Study

Emma Samson[1]*, Philipp E. Hirsch[2,3], Stephen C. F. Palmer[1,4], Jane W. Behrens[5], Tomas Brodin[4] and Justin M. J. Travis[1]

[1] School of Biological Sciences, University of Aberdeen, Aberdeen, United Kingdom, [2] Program Man-Society-Environment, Department of Environmental Sciences, University of Basel, Basel, Switzerland, [3] Research Centre for Sustainable Energy and Water Supply, University of Basel, Basel, Switzerland, [4] Department of Ecology and Environmental Science, Umeå University, Umeå, Sweden, [5] National Institute of Aquatic Resources, Technical University of Denmark, Lyngby, Denmark

OPEN ACCESS

Edited by:
Aristides (Aris) Moustakas,
Universiti Brunei Darussalam, Brunei

Reviewed by:
Slawomir Antoni Lux,
inSilico-IPM, Poland
Antonios D. Mazaris,
Aristotle University of Thessaloniki,
Greece

***Correspondence:**
Emma Samson
emma.sa17@gmail.com

Specialty section:
This article was submitted to
Environmental Informatics,
a section of the journal
Frontiers in Ecology and Evolution

Received: *30 April 2017*
Accepted: *14 November 2017*
Published: *29 November 2017*

Citation:
Samson E, Hirsch PE, Palmer SCF, Behrens JW, Brodin T and Travis JMJ (2017) Early Engagement of Stakeholders with Individual-Based Modeling Can Inform Research for Improving Invasive Species Management: The Round Goby as a Case Study. Front. Ecol. Evol. 5:149. doi: 10.3389/fevo.2017.00149

Individual-based models (IBMs) incorporating realistic representations of key range-front processes such as dispersal can be used as tools to investigate the dynamics of invasive species. Managers can apply insights from these models to take effective action to prevent further spread and prioritize measures preventing establishment of invasive species. We highlight here how early-stage IBMs (constructed under constraints of time and data availability) can also play an important role in defining key research priorities for providing key information on the biology of an invasive species in order that subsequent models can provide robust insight into potential management interventions. The round goby, *Neogobius melanostomus*, is currently spreading through the Baltic Sea, with major negative effects being reported in the wake of its invasion. Together with stakeholders, we parameterize an IBM to investigate the goby's potential spread pattern throughout the Gulf of Gdansk and the Baltic Sea. Model parameters were assigned by integrating information obtained through stakeholder interaction, from scientific literature, or estimated using an inverse modeling approach when not available. IBMs can provide valuable direction to research on invasive species even when there is limited data and/or time available to parameterize/fit them to the degree to which we might aspire in an ideal world. Co-development of models with stakeholders can be used to recognize important invasion patterns, in addition to identifying and estimating unknown environmental parameters, thereby guiding the direction of future research. Well-parameterized and validated models are not required in the earlier stages of the modeling cycle where their main utility is as a tool for thought.

Keywords: individual-based model, RangeShifter, round-goby, invasive, stakeholder, dispersal, management

INTRODUCTION

Invasive Species and the Need for Ecological Modeling

Invasive species are one of the driving forces behind biodiversity loss, and their persistence in non-native areas can result in substantial environmental and economic costs (Pimentel et al., 2000, 2005; Molnar et al., 2008; Cardador et al., 2016). Once established, invasive species have the potential to alter local habitat quality, increase competition for resources, prey on native populations, and spread disease (Kwon et al., 2006; Karlson et al., 2007; Salo et al., 2007; Crowl et al., 2008; Gallardo et al., 2016). As a result, the management and control of invasive species has been a central research focus for many years, and a priority for biological conservation.

There is a continual need for the development and improvement of both new and existing conservation management strategies either to control the spread, reduce biomass or, if possible, to eradicate an invasive species from its non-native environment (Ojaveer et al., 2015). However, implementing management procedures can be costly, both economically and environmentally (Hulme, 2009). Therefore, techniques for forecasting the spread of species and assessing the likely impact of alternative management strategies are desirable (Uden et al., 2015; Kotta et al., 2016; Katsanevakis et al., 2017). One such way to evaluate potential management strategies is through ecological modeling (Uden et al., 2015; Goldstein et al., 2016; Kotta et al., 2016). For example, being able to model the spatial distribution of a species accurately can potentially provide numerous facilities, such as predicting future distributions or furthering our understanding of the original invasion process (Adams et al., 2015).

Forecasting Dispersal in Invasive Species through Spatially Explicit Models

The accuracy and utility of process-based models for ecological forecasting has vastly improved over the past few years (Cuddington et al., 2013; Evans et al., 2013; Urban et al., 2016), particularly as the understanding surrounding ecological processes such as dispersal dynamics has increased (Bocedi et al., 2014; Goldstein et al., 2016). As dispersal is one of the key determinants of species spatial dynamics, understanding and accurately simulating the dispersal process is central to predicting species spread (Hastings et al., 2005; Bocedi et al., 2014; Brown et al., 2014). Numerous studies demonstrate that dispersal is key to species undergoing range expansion, and that there is selection for increased dispersal propensity at the range front (Travis et al., 2010; Brown et al., 2014; Huang et al., 2015; Myles-Gonzalez et al., 2015; Parry et al., 2015; Therry et al., 2015). For example, in the invasive Cane toad (*Rhinella marina* Linnaeus, 1758), individuals in the invasion front disperse further, more frequently and in straighter paths than those in established core populations (Brown et al., 2014; Hudson et al., 2015), and even possess physiology that facilitates their dispersal propensity (Phillips et al., 2006). As such, spatially explicit models that incorporate ecological and even evolutionary or physiological complexity can be vital tools in making predictions regarding range extent and the effectiveness of control regimes for invasive species (Higgins et al., 1996; Meekins and Mccarthy, 2002; Vuilleumier et al., 2011; Goldstein et al., 2016). Calibrating and validating such models with suitable data, if available, can provide an excellent opportunity to investigate species-specific invasions, assess invasion patterns and address concerns. However, very rarely (if ever) will all the data required to parameterize a model fully be readily available in the literature. One way of obtaining such information is through stakeholder interaction. Involving stakeholders in the modeling process additionally allows for the continual evaluation of model utility, accuracy and the development of future model applications.

Early Engagement of Stakeholders in the Ecological Modeling Process

Often stakeholders encounter a model only at the stage where it has been tightly parameterized and validated by ecological researchers. Traditional thinking tends to be that a model needs to be well-parameterized and validated before it can be useful in an applied context. Indeed, an often encountered view is that it can be dangerous for a modeler to demonstrate an "immature" model to stakeholders due to risks of losing credibility or of providing unsound advice. However, developing a well-tested model can be a time consuming process, and this is problematic especially when early intervention is often critical for successful management outcomes. It has been repeatedly highlighted that early involvement of stakeholders into ecological management efforts increases chances for success (Bayliss et al., 2013; Seidl et al., 2013) and we consider that models can provide an important tool for thought at this early stage, well before they reach the level of maturity that we would expect them to have reached prior to providing robust management advice. In assessing the potential risks posed by an invasive species, and scoping out potential control options, scientists and stakeholders must first objectively assess where their knowledge might be incomplete (Krueger et al., 2012) and a prototype model can provide an excellent tool for formalizing the process of establishing what is already known, what is not known and, critically, identifying what it is that isn't known that is likely to be most influential in determining the invasion dynamics. Understanding of where key knowledge gaps exist can inform future research and data collection (Voinov and Bousquet, 2010). Furthermore, the importance of clear communication of the results and implications of risk assessment exercises to stakeholders and authorities has been emphasized, increasingly so in recent years, with particular relevance to the management of our marine ecosystems (Katsanevakis et al., 2017; Stelzenmüller et al., 2018). Here, we put this into practice, and emphasize that it can be extremely valuable to engage stakeholders with an early prototype model and use their input to tailor the modeling process to practical needs. We additionally emphasize the value that an early stage model can provide as a means for horizon scanning for potential threats due to the invasive species, and can be used to provide some initial risk assessments of particular threats (Parrott et al., 2012; Reed et al., 2013; Parrott, 2017).

Case Study: The Round Goby in the Baltic

As a case study, we use our experience of developing an early-stage model for the round goby's spread through the Baltic Sea in order to facilitate stakeholder engagement. The round goby (*Neogobius melanostomus* Pallas, 1811) is a species, for which ecological modeling can be valuable, firstly for formalizing the process of establishing what we know and what we still need to know and, subsequently, for developing well-tested models that can be used to provide robust management recommendations. This species is native to the Ponto-Caspian region, and has invaded the Great Lakes in North America and multiple locations throughout Europe, most likely as a result of transport through shipping routes via ship ballast water (Kornis et al., 2012; Kotta et al., 2016). The species has been termed "one of Europe's 100 worst invaders" and has in a recent evaluation of 18 taxa of non-indigenous species in the Baltic Sea region been found to be amongst those with the greatest impact (Kotta et al., 2016; N'Guyen et al., 2016). For the past 25 years, the species has been in the process of spreading throughout the Baltic Sea (Sapota, 2004; Schrandt et al., 2016). The first reported sighting was in 1990 in the Gulf of Gdansk, and since then, sightings of the species have been recorded in various areas of the Baltic (Kotta et al., 2016). Whilst some stages of the goby's spread have been well-documented, such as the introduction and invasion of the Gulf of Gdansk (Sapota, 2004) and the inner Danish waters (Azour et al., 2015; Carl et al., 2016), there are other stages of the invasion that are substantially lacking in information.

Here, we highlight how a spatially explicit ecological simulation platform, RangeShifter (Bocedi et al., 2014) can rapidly be used to develop an initial prototype model for early engagement of stakeholders with the process, and subsequently calibrated using spatial data available from the literature and input from stakeholders. We then demonstrate how this intermediate-stage model can be applied to further research in order to identify key data gaps that would need to be filled before a well-tested model could be used to robustly inform management actions.

OVERVIEW OF THE PROCESS

The work described in this paper has been designed to be consistent with the adaptive modeling approach for ecological forecasting outlined in Urban et al. (2016). The overall process of developing the model is outlined in **Figure 1**. A prototype model of the goby's spread throughout the Baltic was developed and parameterized within a 6 week period through an iterative process (Grimm et al., 2005; Grimm and Railsback, 2012) to present to stakeholders in a symposium context. This period of initial model development was by necessity short in our case, as we had been invited to a round goby symposium to discuss the potential utility of the RangeShifter software in the context of managing the round goby. The description of this initial model development will be kept brief, as it was predominantly an iterative process of altering parameter values and comparing the model output to that of the HELCOM round goby distribution (Michalek et al., 2012). The rapid production of a prototype model allowed demonstration of the potential utility of the model to stakeholders, especially for use in the future after more rigorous assessment. Furthermore, it also provided an overview of what the model could do, which opened the way for suggestions on scenarios and improvements that the model can be used to explore.

Stakeholder Collaboration

Overview of the Symposium

A symposium centered around the spread and impact of the round goby in the Baltic Sea was held in Kalmar, Sweden in June 2016. The organization of the symposium was headed by the Swedish Agency for Marine and Freshwater Management[1], and there were an estimated 30 attendees. The main stakeholder groups consisted of representatives of different levels of local and regional environmental administration, people that participated in a private capacity, and representatives of other groups interested or affected by round goby spread, such as recreational and professional fishermen. The symposium was followed by a workshop focusing on solutions to manage and impede the spread of the round goby throughout the Baltic Sea. During the symposium the overall research project and the model was presented in a 30 min power point presentation (Supplementary Figure 1). The presentation had two main components. First, RangeShifter was presented to the participants along with examples of how the software had already been used to address conservation relevant questions, including invasive species. This was key as a means for establishing our credibility as modelers. Second, the prototype goby model, implemented using RangeShifter, was presented to the stakeholders to demonstrate the potential utility of the model within the Baltic Sea and hence within the geographical focus of the participants' interests. Throughout this second part, we repeatedly stressed both the prototype nature of the model and the fact that while we were in a room full of goby experts, the modelers who had rapidly developed a prototype for demonstration were certainly not.

At the end of the presentation a specific call for input was issued: a slide stating "What we hope to get from you…" followed by six suggested inputs: Specific parameters (e.g., demographic and dispersal), The estimated introduction sites (and when), Patterns for comparing model outputs with spatial and temporal patterns of density and sediment type and habitat, Proposed management techniques. Following the presentation there was an open discussion with a call for feedback and input. In transdisciplinary projects it is important that both scientists and non-academic partners contribute on an equal footing (Hadorn et al., 2008; N'Guyen et al., 2016). This is especially relevant in the case when inputs are qualitative rather than quantitative. In our case, we were interested in qualitative inputs, and we therefore designed the interaction with stakeholders as open and did not follow a standardized procedure. We felt this would ensure an

---

[1] https://www.havochvatten.se/

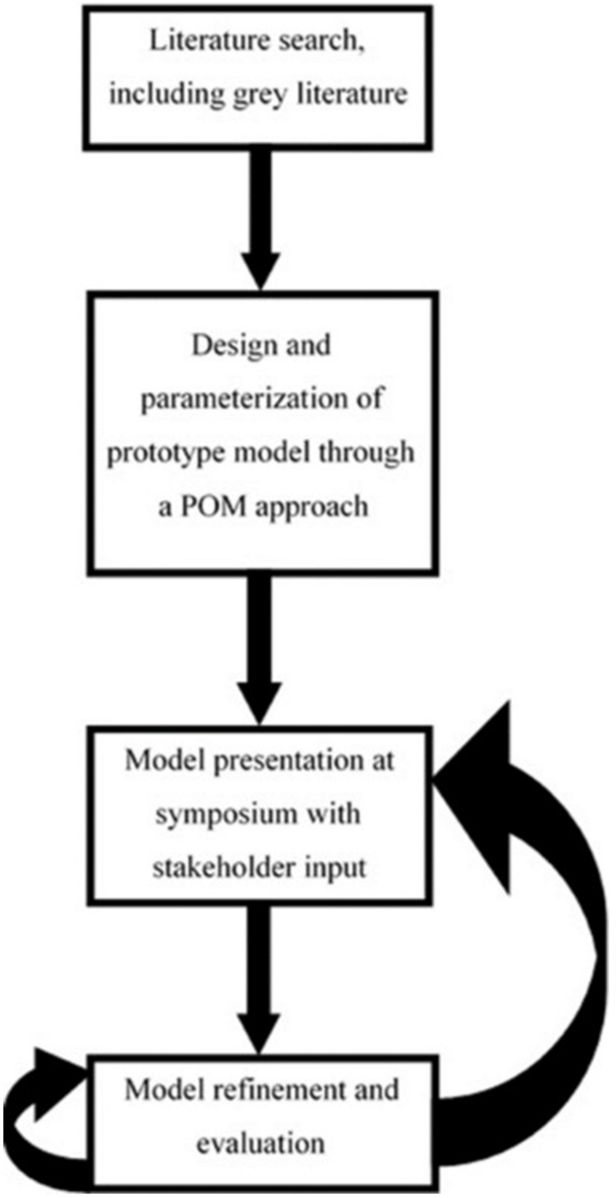


FIGURE 1 | A schematic representation of the modeling process, from initial literature search to the proposed next steps. Model refinement and evaluation is iterative, reflecting the alterations that are constantly made to the model during the calibration process. Once refined, this model should then be reintroduced to stakeholders for further co-development.

atmosphere that encouraged stakeholders to contribute even anecdotal but possibly relevant information which they might be less inclined to share when e.g., filling out a questionnaire.

Outcomes of the Symposium

The interaction with stakeholders identified essential knowledge gaps, which would have gone unnoticed by us as scientists alone. Crucially the interactions also provided a clear focus in terms of what a useful model would need to include and would need to be able to predict in order that it was most useful to the stakeholders. Also, personal communications with multiple researchers and stakeholders present at the meeting provided an insight into the current understanding of the round goby's spatial presence in the Baltic Sea that was not obvious from searching the literature, including information on new studies that will yield high quality data. Three essential qualitative outcomes of the symposium that were derived from the interactions between modeling team and stakeholders provided strong focus for future work. These related to model building such that key processes driving the spread dynamic are properly represented and parameterized and to developing the model to ensure its relevance for informing key management decisions:

First: A knowledge gap regarding the depth of goby dispersal was highlighted as potentially crucial. Prototype model results shown at the workshop included one suggesting that the invasion dynamic is likely to be very sensitive to the depth range over which gobies can disperse. At the workshop attendees noted that adult gobies are sometime caught in deeper water. However, it was suggested that this occurs during winter months and may reflect some adults exhibiting seasonal migration to deeper waters. It became obvious that whether gobies disperse through deep water or disperse solely in shallow areas is currently unknown. Understanding the depth range of goby dispersal may be of great importance to those involved with the round goby invasion for a number of reasons. Depth acting as a barrier to dispersal may be utilized in numerous management protocols to impede or inhibit goby spread into undesirable areas. Furthermore, understanding goby dispersal depth helps to predict future areas that may be under threat of round goby invasion, even without a human-mediated element to the dispersal. Identifying the potential importance of the depth sensitivity of dispersal for patterns of goby spread was a novel outcome of the workshop that will motivate new empirical work.

Second: Threats of the round goby's invasion of the freshwater systems that connect to the Baltic Sea, particularly with regards to Salmonids were identified, as the round goby may devastate their populations through egg consumption (Chotkowski and Ellen Marsden, 1999; Marentette et al., 2011; Ladago et al., 2016). This potential impact of the round goby was a key issue for many of the stakeholders present and highlighted the importance that to be useful for management a model would need to be able to effectively operate into riverine systems and potentially also account for salinity gradients and tolerances.

Third; The threat that the round goby poses to the long-tailed duck (*Clangula hyemalis* Linnaeus, 1758) was emphasized (Hearn et al., 2015). The Baltic Sea is the key wintering destination for the majority of the western Siberian and northern European populations of the long-tailed duck (Hearn et al., 2015), which currently faces a multitude of threats such as predation, competition, oil spills, gillnets, hunting, habitat destruction, and water traffic (information available on the BirdLife International website[2]). The round goby and the long-tailed duck share a diet of mussels and crustaceans. Hence, the spread of the round goby to the overwintering habitats may result in competition for food. As the Baltic Sea is the main overwintering area, a reduction

---

[2] http://www.birdlife.org

in food availability for the long-tailed duck in this area may prove devastating. Consequently, this area is recommended to be a crucial area to protect from invasion by the round goby (Hearn et al., 2015). Currently there are no effective means for estimating the risk that these areas will be invaded. Hence, estimations for whether, and if so when, the goby will reach the overwintering areas from their current distribution would be valuable, to estimate time-scales for conservation efforts for the long-tailed duck, and to design measures to protect the area from the spread of the round goby. We note here that while there was existing information in the literature highlighting the potential impact of round goby on long tailed duck in the duck's key overwintering sites (Hearn et al., 2015), it would have been unlikely that the modeling team would have easily found it. Thus, the stakeholder workshop provided a means for those fully familiar with the system to direct the modeling team to literature relating to the focal species and its potential impacts that isn't primarily about the focal species.

The Modeling

Modeling Population Dynamics and Dispersal in RangeShifter

We used a spatially explicit, individual-based model (IBM), RangeShifter (Bocedi et al., 2014) to simulate the spread of the round goby throughout the Baltic Sea. RangeShifter was developed in response to the demand for integrated dynamic models, and as such, provides a platform with which to model complex population dynamics and dispersal behaviors, at the individual scale (Franklin, 2010; Huntley et al., 2010; Thuiller et al., 2013; Lurgi et al., 2015).

To represent the Baltic Sea, a gridded seascape was created in ArcGIS 10.3.1 using raster data extracted from the EMODnet Bathymetry portal[3] Each cell was 2.5 by 2.5 km and characterized by depth. Population dynamics were modeled at the cell scale. The numbers of individual fish in the Baltic, or even in a local area, at reported densities (Vélez-Espino et al., 2010) would be far too large to be explicitly represented in the model, and therefore we treated a modeled "individual" as representing a localized established sub-population of unspecified size (hereafter "individual" for consistency with RangeShifter terminology), which was regarded as female in a single-sex model. It was not necessary to represent the overlapping generations of the species, but sufficient to model the reproductive rate of such "individuals," i.e., the rate at which "daughter" sub-populations were produced, some of which would disperse to expand the range of the species.

At model initiation, individuals were assigned to cells within species introduction points at half carrying capacity. In each year, the overall dynamics consists of reproduction, death of adults, and offspring dispersal. Reproduction by each individual is determined by a stochastic draw from a Poisson distribution having a mean set by the maximum growth rate at low density and subject to density-dependent reduction following Maynard-Smith and Slatkin (1973). Carrying capacity, *K*, was set to 10 individuals/ha for all cells. However, this limitation is unlikely to be critical for the pattern of range expansion on which we were focused, given that densities at the range front are expected to be much lower than in long-established areas (Brownscombe et al., 2012; Groen et al., 2012; Azour et al., 2015).

Once reproduction has taken place, individuals could emigrate away from their natal cell, an action dependent on the local density within the cell. If an individual left the cell, its trajectory was modeled using the Stochastic Movement Simulator (SMS; Palmer et al., 2011). SMS simulates an individual's path throughout the landscape, in which the direction of movement between cells is based on the relative cell "costs" to movement and on a tendency to follow a correlated path (directional persistence). The perceptual range, in which costs were evaluated, was set at 1 cell (i.e., no more than 2.5 km).

Incorporating the Stakeholder Input into the Model

A key issue that emerged from the stakeholder workshop was a lack of knowledge relating to the depths of water through which gobies can disperse. This issue was, in part, highlighted by some of the runs of the prototype model, demonstrated at the workshop, in which it was clear that including a depth threshold resulted in very different spread patterns than omitting one. Accordingly, cell cost was set in relation to a threshold depth for movement: the cost of traversing a cell of the depth threshold and deeper was set to a very high value, and the cost of traversing a cell above the depth threshold was set to a very low value. In doing so, individuals were much less likely to travel into deeper water than that set by the threshold. For all depths, each step an individual took had an associated spatially and temporally constant mortality risk.

Upon reaching a new cell, an individual had the opportunity to settle or continue movement to a different cell. The decision to settle was density-dependent. If the population density was too high in a cell, then the individual would not settle but continue to disperse to a neighboring cell (Bocedi et al., 2014).

Parameter Calibration and Assessing Model Performance

The majority of the parameters required for the model were not widely available in the literature or through online resources. Consequently, in order to calibrate the model parameters, the Gulf of Gdansk was chosen, as detailed spatial information regarding the goby's spread through the area was available. This spatial information was primarily obtained from the NOBANIS fact sheet, produced by Sapota (2012). NOBANIS is the European Network on Invasive Alien Species, and the project produces information and fact sheets on invasive alien species. The fact sheet, written and referenced by experts, provides a range of information including recommendations for management, species ecology and information regarding its historical introduction and spread. This temporal spatial presence information available in the NOBANIS fact sheet was used as a baseline to calibrate the model.

Parameter values were calibrated using a pattern-oriented modeling approach (POM) (Grimm et al., 2005; Grimm and Railsback, 2012; Bauduin et al., 2016), in which simulations

---

[3] http://www.emodnet-bathymetry.eu/

were run for a variety of values for four key parameters, namely the maximum growth rate, the depth threshold, the per-step mortality risk, and the maximum settlement probability (**Table 1**), in order to find a combination which most precisely matched that of the historical round goby spread throughout the Gulf of Gdansk. For each simulation, the final model distribution was compared to the actual distribution reported in Sapota (2012). Other more minor parameter values, such as the depth threshold cost, were assigned during the creation of the prototype model, using an iterative process. During this process, the values chosen were arbitrary, and altered until the model output started to match the goby distribution seen in the NOBANIS fact sheet. Therefore, these parameters were used more as values to tune the initial model, rather than parameters that were important to investigate. The model's predicted output was compared to the observed output for each year that data were available, in order to obtain the most accurate dispersal pattern throughout the Gulf of Gdansk.

TABLE 1 | RangeShifter settings and parameter values for Gulf of Gdansk and Baltic Sea models.

Parameter	Description	Gdansk	Baltic
	Cell-based landscape, cell size	2,500 m	2,500 m
	Rows × Columns	48 × 43	625 × 717
	Habitat codes (representing depth classes)	1–12	1–12
	Female-only model, no stage structure		
K	Carrying capacity (per ha) (all habitats)	10.0	10.0
Rmax**	Mean growth rate at low density	1.2, 1.4, 1.6	1.2, 1.4, 1.6
bc	Competition coefficient	1.0	1.0
d	Density-independent emigration rate	0.7	0.7
	Transfer model–SMS		
	Cost for depth layers above threshold	1	1
	Cost for depth layers below threshold	100,000	100,000
PR	Perceptual range (cells)	1	1
PRmethod	Perceptual range method	1	1
DP	Directional persistence	1.0	1.0
SMconst**	Per-step mortality risk	0.1, 0.2, 0.3, 0.4	0.1, 0.2, 0.3, 0.4
	Density-dependent settlement:		
S0**	Maximum probability	0.4, 0.5, 0.6, 0.7, 0.8, 0.9	0.4, 0.5, 0.6, 0.7, 0.8
alphaS	Slope	−10.0	−10.0
betaS	Inflection point	1.1	1.1

***For the Gulf of Gdansk, three levels of Rmax, four levels of SMconst, six levels of S0, and 11 depth thresholds were applied in a fully factorial design yielding 792 simulations, each of which was replicated 100 times. For the Baltic Sea, a partially factorial set of 48 combinations of Rmax, SMconst, S0, and four depth thresholds (selected from the Gulf of Gdansk model) were each replicated 100 times.*

Accuracy of Model Calibration

To assess the accuracy of the model for each parameter combination, four metrics were used, in addition to visually inspecting the model output. Model *specificity* (in which both the observed distribution and the model's predicted distribution do not have individuals present in a cell), *sensitivity* (in which both the actual distribution and the model have individuals present in a cell), the receiver operator characteristic (ROC) curve with the associated area under the curve (AUC), and Cohen's Kappa, κ. The κ statistic represents a way to measure reliability, or precision, and compares the model prediction accuracy with the accuracy expected to occur by chance (Allouche et al., 2006). Sensitivity and specificity both vary from −1 to +1, in which a score of 0 represents no better than chance, and +1 would represent a perfect score. κ can vary between 0 and 1, where 0 represents an agreement no better than chance, and 1 represents a perfect agreement. An accurate model with an AUC score of an ROC plot would be close to 1. A score close to 0.5 would represent a poor model. Whilst the AUC is threshold independent, the other measurements are threshold dependent. The threshold during analysis is the cut-off value used to translate predicted probabilities into a presence or an absence. Consequently, for a predicted probability to be classed as a presence under a high threshold (such as 0.9), a cell would need to be colonized by individuals in 90% or more of replicated model runs given the specified combination of model parameters.

In order to calculate the sensitivity, specificity, AUC and κ for each parameter combination, each combination was repeated over 100 simulations. These metrics were calculated using the PresenceAbsence package in RStudio 3.3.1 (Freeman and Moisen, 2008).

Ports of Introduction and Modeled Population Initiation in the Gulf of Gdansk and the Baltic Sea

The species introduction points of the Baltic Sea, where populations were initiated, were estimated using information available in the literature (Kotta et al., 2016), species presence information from the symposium and shipping port and traffic information available on Baltic Transport Maps[4] Ports of introduction were assumed to be the closest shipping port to a current goby distribution. The initiation of a population at the entry points was staggered in an attempt to replicate the introduction of the goby throughout the Baltic at various points in time. For example, populations were initiated in the Gulf of Gdansk entry points at year 0 (representing the year 1990), but populations initiated around Kalmar were not initiated until year 20 (2010). The timing of the staggered introductions at various points on the map were based on estimates from the literature (reviewed by Kotta et al., 2016). The staggered introductions were carried out using a customized version of RangeShifter that allowed populations to be initialized in individual cells at specified times. Parameter values applied were informed by the results from fitting the model to the Gulf of Gdansk (**Table 1**).

---

[4] http://www.baltictransportmaps.com

MODEL RESULTS

Model Calibration: Role of Depth Threshold in the Gulf of Gdansk

Model accuracy was most strongly influenced by the depth threshold: 76% of the variance in κ was explained by depth, as compared with 11% by the maximum settlement probability, 4.6% by the per-step mortality risk, 3.3% by the maximum growth rate and negligible amounts by interactions. The model was most accurate for a depth threshold between 10 and 25 m, and accuracy increased slightly with decreasing settlement probability and mortality risk and with increasing growth rate (**Figure 2**). Similar conclusions regarding the importance of the depth threshold were drawn from the other accuracy metrics (**Table 2**; Supplementary Table 1). Examples of various outputs can be seen in **Figure 3**, ranging from good (AUC and $_k$ scores close to 1) to poor (AUC and $_k$ scores close to 0.5 and 0, respectively). Through the process, a number of models with high accuracy were produced, with some models obtaining accuracy values of 0.8 for all four accuracy metrics, even when the model threshold was high (0.8) (**Figure 4**).

Model Output: Projections Based on the Role of Depth Threshold, across the Entire Baltic Sea

Despite obtaining a range of accurate parameter combinations for the Gulf of Gdansk, when they were applied to the entire Baltic, the overall model output was poor when compared to the extensive observed distribution spanning a substantial proportion of the Baltic coastline as reported in the literature (**Figure 5**). The accuracy scores calculated for The Baltic suggest that the model was not much better at predicting the goby distribution than chance (AUC scores close to 0.5, and other scores close to 0.1).

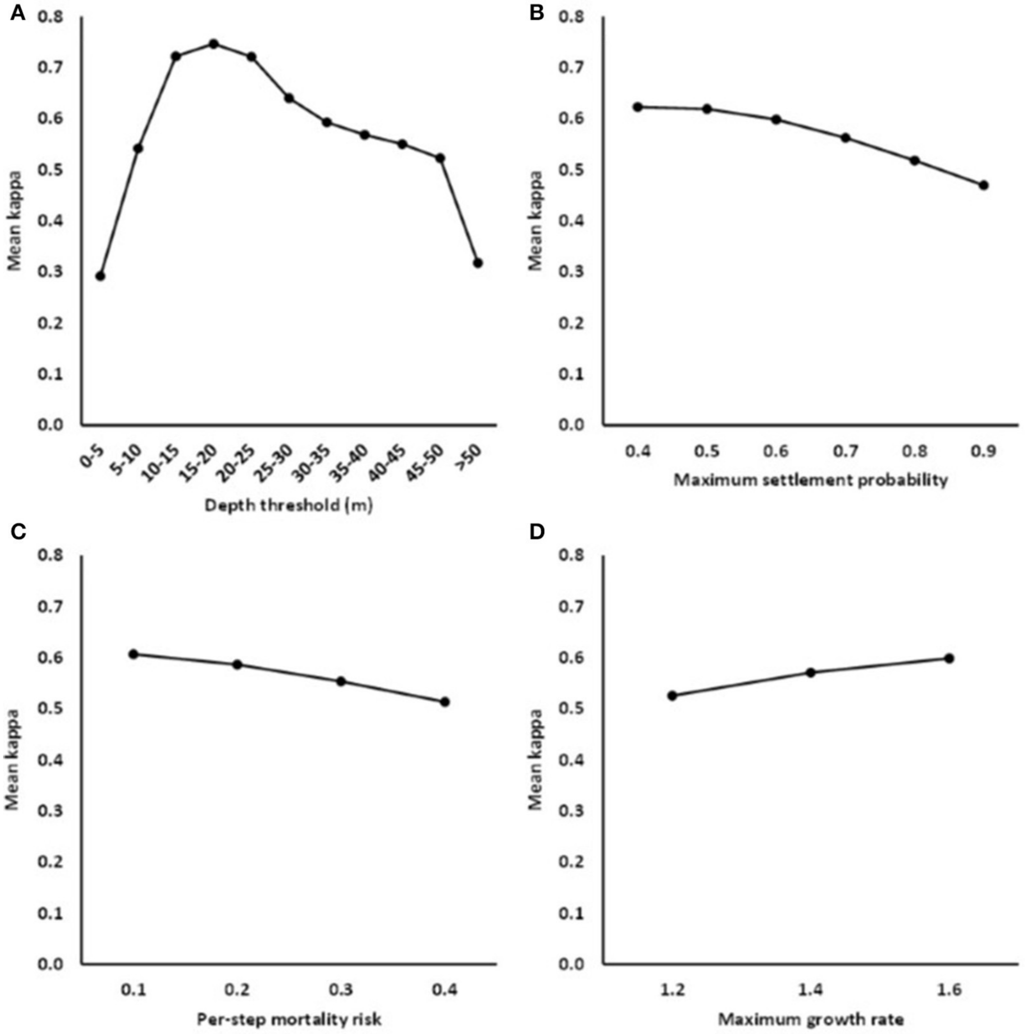


FIGURE 2 | Fit of the RangeShifter model for the Gulf of Gdansk: marginal mean values of κ (kappa) in relation to **(A)** the depth threshold class, **(B)** the maximum settlement probability, **(C)** the risk of mortality at each step taken during dispersal, and **(D)** the maximum population growth rate.

TABLE 2 | An example of the effect of varying the depth threshold of dispersal on the accuracy of the predicted population distribution, all other parameters being held constant.

Depth threshold (m)	Sensitivity	Specificity	Kappa	AUC
0–5	0.587	0.876	0.432	0.843
5–10	0.471	0.997	0.588	0.803
10–15	0.740	0.993	0.807	0.906
15–20	**0.888**	**0.976**	**0.862**	**0.959**
20–25	**0.915**	**0.964**	**0.848**	**0.970**
25–30	0.897	0.930	0.757	0.957
30–35	0.892	0.904	0.699	0.944
35–40	0.888	0.889	0.666	0.935
40–45	0.883	0.881	0.647	0.932
45–50	0.883	0.871	0.628	0.924
Below 50	0.901	0.689	0.379	0.836

For all three model assessment parameters, values over 0.8 represent a highly accurate model fit. The model threshold for the cut-off (i.e., above which the predicted probability was regarded as presence, and below which as absence) was 0.8. The most accurate models are displayed in bold text.

DISCUSSION

In this study, we rapidly developed a prototype model of round goby spatial dynamics that was used to facilitate early engagement with stakeholders. We subsequently combined data available in the literature and stakeholder input in order to calibrate the IBM such that it simulated the round goby's spread throughout the Gulf of Gdansk to a high level of accuracy. We then used the calibrated model to simulate its spread through the Baltic Sea, despite the limitation of imprecise and potentially inaccurate presence data. Our experience demonstrates the value of involving stakeholders early in the modeling process. Prototype model results had indicated that predicted spread was highly sensitive to the inclusion of a depth threshold for dispersal, and the subsequent stakeholder communication highlighted how little is currently understood about goby dispersal at various depths. Consequently, various depth thresholds were incorporated into the modeling, in order to assess the impact of depth on model accuracy and therefore goby dispersal. We demonstrated how, by using known spread patterns, it can be possible to use the model to infer details of the dispersal process, in this case related to the depth threshold of goby dispersal. In detail, we could show that that the limit to dispersal depth of the round goby lies between 10 and 25 m. Empirical data are now required on the depth sensitivity of dispersal such that a robustly parameterized model can be used by the stakeholder/modeler grouping in further steps toward identifying management options. The involvement of stakeholders as early as possible in the process and their regular inclusion throughout as co-developers of the modeling will facilitate a cooperation between scientists and stakeholders in putting possible management measures into practice.

Stakeholder Collaboration—Putting Theory to Practice

Research has identified that the long lag time between research and its publication hinders managers of biological invasions to make use of important results such as our models generated (Matzek et al., 2015). In addition, theory predicts that the success rate of management should be higher if stakeholders and scientists engage early on in the transdisciplinary process of managing an invasive species (Hirsch et al., 2016a; N'Guyen et al., 2016). The main reason behind this is that scientific results that were co-produced by relevant parties in a transdisciplinary process should have better social acceptance and higher compliance by decision makers (Pohl and Hadorn, 2008).

In our study, we put these theoretical predictions into practice and engaged in a modeling process that used stakeholder input as an essential component. Stakeholders provided two essential inputs regarding future model optimization: providing information on where higher quality distribution data would be available in the near future and on the priority of including depth in the model. Stakeholders contributed their knowledge and understanding on an equal footing. In an excellent recent contribution on how to co-develop models with stakeholders effectively to address pressing ecological problems, Parrott (2017) argues that it is important for the modelers to get to know the study system well before meeting with stakeholders. Parrott (2017) writes, "Knowing the system well is a key to gaining the trust and confidence of stakeholders in the ability of the modeler and the entire research team to contribute meaningfully to the issue. If the researchers are not from the area, they should spend time visiting and getting to know the region before initial meetings with stakeholders." We had been approached by stakeholders and asked to present the modeling software at a meeting on the threat posed by round gobies to illustrate what might be possible in terms of using RangeShifter to inform management of the species. We only had a few weeks ahead of the meeting in which we were able to build a prototype model for the goby and were thus unable to acquire substantial knowledge of the system prior to meeting stakeholders. However, at the meeting we were able to demonstrate our credibility as ecological modelers by first providing examples of how the RangeShifter was being used to address a range of other applied issues, including landscape management to conserve African forest birds (Aben et al., 2016), assisted colonization of butterflies in Finland (Heikkinen et al., 2014) and the invasion of American mink (*Neovison vison*) in Scotland (Fraser et al., 2015).

Acknowledging the Different Roles of Scientists and Stakeholders

A potential advantage of the approach we took in this study is that the stakeholders naturally take the role as the species/system experts, and the potential risk whereby stakeholders perceive that the researchers assume the role of experts and tell them how their system works is reduced. One potential disadvantage of such an approach is that researchers cannot glean data from stakeholders in the form of quantitative assessments through e.g., specifically designed questionnaires. This disadvantage, however, is compensated by the fact that stakeholders can contribute their knowledge freely through unstructured interactions with researchers. For that, it is clearly critical that the modeling team gain the confidence of the stakeholders, but that need not be by having acquired detailed understanding of the particular study system in advance of a first meeting. Indeed, we suggest that the effective establishment of a model co-development group may be

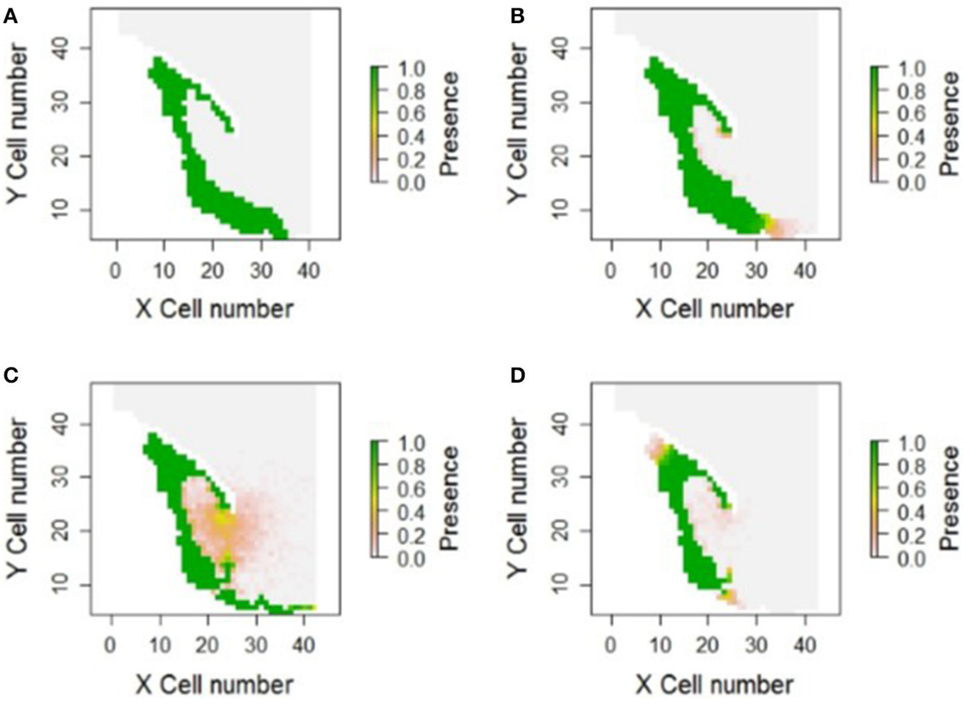


FIGURE 3 | Example model outputs from four different parameter combinations in the Gulf of Gdansk. X and Y cell numbers represent the cell number, or coordinates, on the gridded landscape created for the modeling exercise. Green cells represent a cell that was colonized by populations in each of the 100 repetitions (i.e., 1.0 refers to 100% of repetitions). Model **(A)** represents the actual goby distribution, and therefore a perfect model output. Model **(B)** represents an example of an accurate model, whereas **(C)** represents model over dispersion, and **(D)** under dispersion combined with dispersion into the wrong depths.

facilitated if this is actually not the case and at the start of the process there is a clear division of expertise between modelers and stakeholders. As the process of co-development of a model proceeds, both researchers and stakeholders can build upon this first interaction on an equal footing albeit with quite different expertise. Our study provides a practical example for future model building efforts on how to rapidly initiate transdisciplinary projects, which is absolutely vital if models are to be successfully used to inform early intervention against invasive species.

Model Calibration

Calibrating the model with precise spatial data produced a highly accurate model that simulated the spread of the goby throughout the Gulf of Gdansk over an 11 year period. The model outputs obtained from the calibration process highlighted the key role of the depth threshold to movement. However, when scaled up through space and applied to the whole of Baltic Sea, the model failed to predict a distribution similar to that observed in the literature. The failure to produce a model for the Baltic Sea with a high degree of accuracy has several implications.

One of the main downfalls of the Baltic model seems to occur from uncertainty regarding introduction points. In order to obtain a predicted presence from the model that was similar to that of the observed presence, further introduction points would need to be added, if the parameters obtained from Gdansk were to be used. Although short-distance (∼30 km/year) active migration appears to occur in some local areas (Azour et al., 2015), this suggests that, at the scale of the Baltic sea, the goby did not disperse over long distances as a primary mode of invasion, but that human-mediated transport, for example via ships or other means, was the primary cause of invasion. As large ports were used in the model as the introduction points, this may also suggest that the goby was introduced to various areas that were not necessarily large commercial ports, but also small recreational ports. Subsequently, future efforts to manage the spread of the goby may benefit from focusing preventative measures on human-mediated transport, such as the cleaning of recreational boats (Hirsch et al., 2016c). This will be particularly important in protecting regions that would otherwise be likely to be out of the range of goby colonization due to their being effectively isolated by channels of deeper water.

Furthermore, the presence data used to produce the observed map for model calibration was at a coarse spatial scale. It may be that the goby's presence at various depths in the Baltic was not represented in the observed distribution at a fine enough resolution for accurate model assessment. Given more precise presence data, at a finer resolution, the accuracy of the models predicted goby presence in the Baltic Sea could improve substantially. One of the benefits of such models is the ability to identify on which future data collection efforts should be focused. This is in agreement with the recent call for mandatory catch records and citizen science programs in order to collect data on the round goby (Ojaveer et al., 2015). In the case of this modeling

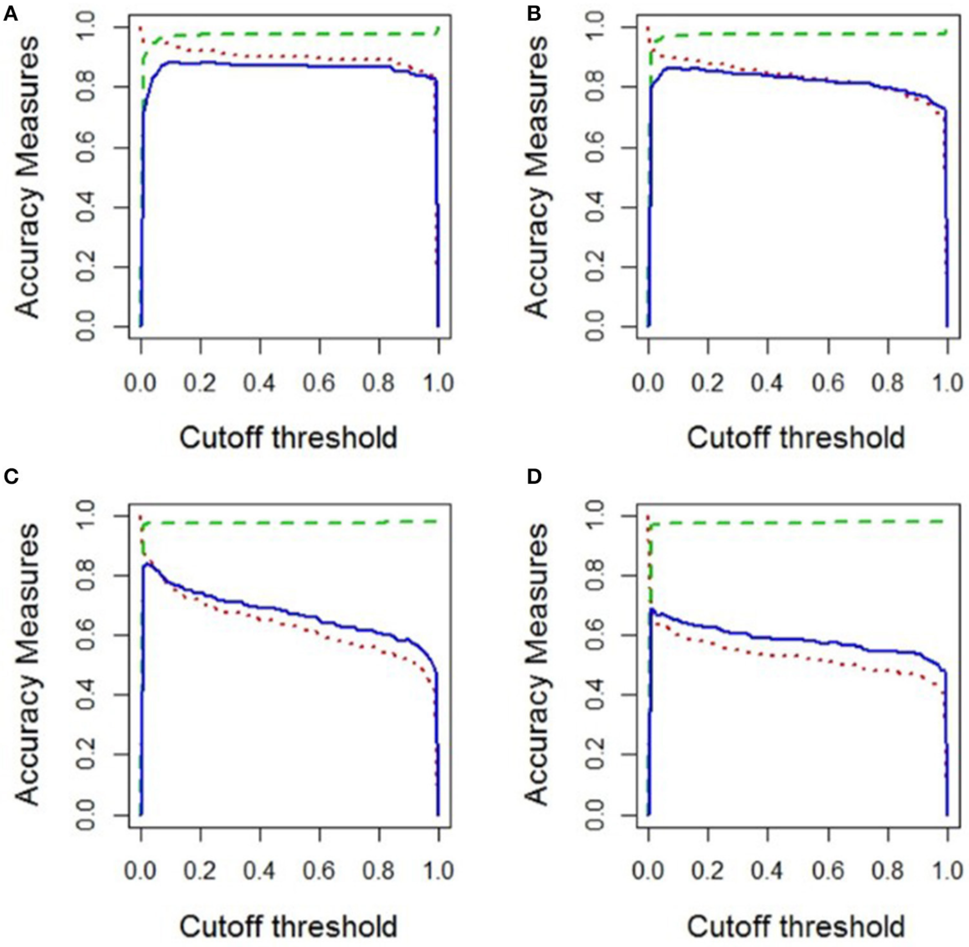


FIGURE 4 | An example of the plots used to assess the accuracy of parameter combinations, using kappa, specificity, and sensitivity measures. The accuracy measures vary from zero to one, in which a value of one represents a perfect accuracy measure and zero a poor one. The cutoff threshold represents the number of repeat simulations a cell was required to have been colonized, in order to be characterized as a presence in the final model evaluation, with 1.0 being 100% and 0.0 being 0% of repeats. **(A)** represents a set of parameter combinations that predict a goby presence close to that of the observed presence from the literature. In comparison, plots **(B–D)** demonstrate a decrease in model accuracy.

exercise, presence data over various depth distributions, and the identification and incorporation of the correct introduction points, have been identified as being critical for accurate model calibration.

Depth Sensitivity

In order to replicate the observed goby distribution throughout the Gulf of Gdansk, a dispersal depth limit of ∼20 ± 5 m produced the most accurate model. It is nevertheless important to note that this was calibrated using one area of the Baltic Sea. Thus, to obtain more accurate results, presence data spanning various depths over more locations in the Baltic Sea are required. Hitherto there have however not been any studies dedicated to investigating this aspect of the biology of the species. Furthermore, as round goby is not a commercial species, no catch-related depth information is available from the fishery. The sparse information that exists is from a Polish young fish surveys program, showing that, although generally considered a shallow water inhabitant, high catch rates occur at 50–60 m depth during winter months (November and January–March) (Grygiel, 2007). This suggests that during the cold season, the fish is wintering in deeper sea areas, but whether dispersal occurs during this period or when the fish resides in more shallow, coastal waters remains speculative. The present modeling exercise thus indicates that future research efforts should prioritize obtaining presence and absence data for round goby at various depths throughout the Baltic, and investigate whether dispersal to novel areas occurs during the warm or cold season. Although often expensive and time consuming to collect, this type of information has been achieved for several species though tagging studies (e.g., Boje et al., 2014). Furthermore, compilation of existing data from various national and international surveys and monitoring programs (e.g., the biannual Baltic International Trawl Surveys, BITS) could prove to be a cost-efficient way to obtain essential information. The depth threshold of round goby dispersal is an essential parameter not only for calibrating models, but also for

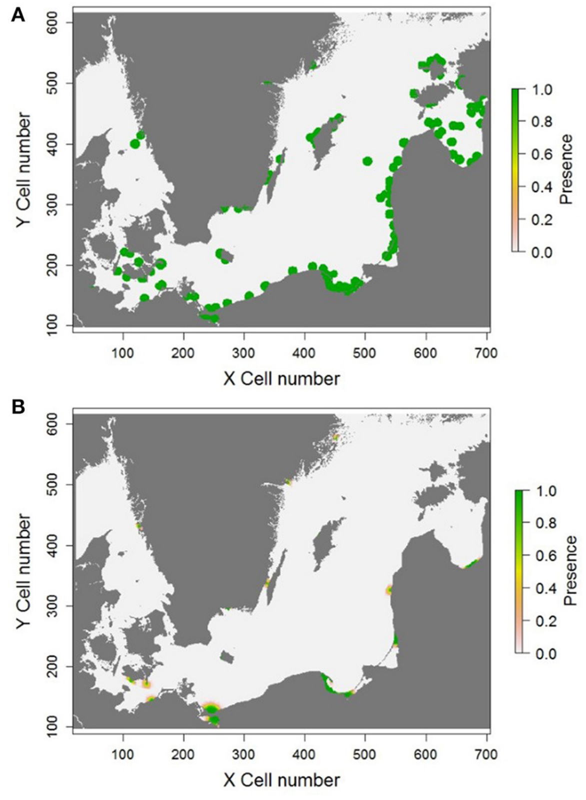


FIGURE 5 | Comparison between **(A)**, the observed goby distribution available from the literature, and **(B)**, an example predicted distribution obtained from the model. The color of the cell represents the presence of a population in a cell, and therefore its colonization in a repetition. The presence varies from one to zero, with a value of one meaning the cell was colonized in every repetition and a value of zero meaning the cell was never colonized. The values along the axis represent the cell numbers of the landscape grid used in the modeling exercise, in which each cell size was 2.5 km.

incorporating into risk assessments of the species spread, both generally and for areas of special interest.

Salinity Tolerance and Ecological Parameters Influencing Spread

Although not identified by the stakeholders in the present study, parameters besides depth should be evaluated for their potential relevance for dispersal tendency. Charlebois et al. (2001) highlighted the need for research determining "dispersal mechanisms and environmental characteristics that limit dispersal." Round goby is considered a euryhaline species, which is able to adapt to salinities ranging from freshwater to brackish conditions. Previous studies have suggested that round goby will not endure oceanic conditions (i.e., high salinity) (Ellis and Macisaac, 2008; Karsiotis et al., 2012). A recent study acclimating round goby to salinities spanning from fresh to seawater has shown that slow increases in salinity (5 PSU per week) to salinities approaching oceanic conditions (30 PSU) severely affected the osmoregulatory capacity of round goby. Although survival was also reduced at oceanic salinities, still 61% of the fish survived at 30 PSU. So while salinity will likely not act as an effective barrier, it might still impede the ongoing spread of round goby through the salinity gradient from the brackish Baltic Sea and into the oceanic North Sea and this warrants its inclusion into dispersal models (Behrens et al., 2017). Further parameters which could turn out to be relevant depend on the study system and could include temperature (thermal limits in round goby are between 0.5 and 26°C (Chekunova 1974 cited in Charlesbois et al., 2001) and, in running waters, flow velocity (round goby show a critical swimming speed of 35.5 cm s^{-1}; Tierney et al., 2011). Recent research suggests that population niche modeling in combination with climatic parameters might benefit from the introduction of thresholds for certain environmental parameters (Almpanidou et al., 2016). Incorporating a minimum of climatic suitability might allow coupling of dispersal models with models of population establishment (Almpanidou et al., 2016). Understanding the interplay of population dynamics and dispersal is relevant for selecting population management options in newly identified populations (N'Guyen et al., 2017).

Personality-Dependent Behavior as a Model Parameter and Management Options

Not only the abiotic environment, but also personality-dependent behaviors can be important at the invasion front, where local sub-populations consist mostly of bigger/older asocial individuals (Thorlacius et al., 2015). Recent research has found that personality traits can inform models of dispersal such that only individuals showing trait values above a certain threshold are predicted to disperse (Hirsch et al., 2017). In combination with the depth thresholds, such an approach can complement future models to achieve an even higher accuracy in predicting dispersal.

Until further information is available, our modeled depth trial results may be used as a preliminary guide to assess management regimes and prioritize management areas for vulnerable species. For example, from an applied perspective, the model results raise the prospect that artificial deep channels may stymie the spread of the species. Telemetry-based data on the spread of invasive crayfish in a Central European large lake has also suggested a spread along the shoreline down toward a certain depth isocline. This might make it plausible to slow the natural spread by barriers (Hirsch et al., 2016b). In Lake Tahoe, USA, invasive bivalves have been successfully controlled by the installation of gas impermeable benthic barriers (Wittmann et al., 2012). These examples demonstrate how knowledge of the spatial spread of an invasive species can directly inform its management.

Practical Model Application for Protecting the Long-Tailed Duck

In a practical application of this transdisciplinary approach, we designed a preliminary modeling experiment as an example of how detailed models developed with stakeholders can inform risk assessment of invasive species and help to identify priority areas for management. A key issue that emerged through the stakeholder interaction is the implication of the goby's invasion of the over-wintering habitat of the western Siberian and northern European wintering populations of the endangered long-tailed duck. The populations of the duck may be threatened by the round goby through exploitative competition for food (Hearn et al., 2015; Skabeikis and Lesutiene, 2015; BirdLife International website). In a preliminary trial, we used the calibrated model to demonstrate how an effectively parameterized model could be used to assess whether the long-tailed duck overwintering habitat was at risk of colonization from the round goby in the future. Again, we did this by running the model over a number of years and a number of depth thresholds. This produced a number of scenarios in which the overwintering habitat of the long-tailed duck was invaded, but the time it took for the invasion depended greatly on the depth threshold at which the round goby was able to disperse. Given the current uncertainty surrounding the results of these initial trials or the risk to long-tailed duck populations, and the potential influence results from them might have, we decided that it was premature to publish the results at this stage. However, whilst only being a preliminary experiment, this example reinforces generally (and very effectively reinforced to our co-development team of modelers and stakeholders) the importance of obtaining accurate spatial data regarding the presence of the round goby at various depths.

Future Modeling Perspectives

In this study, we made use of an IBM to simulate the spread of the invasive species. However, it is important to recognize that alternative approaches exist that could equally well be used in transdisciplinary work where models are co-developed to inform understanding and management of invasive species. Indeed, in future studies one valuable approach will be to utilize more than one of these modeling approaches in concert. For example, there can be considerable benefits of jointly developing a stochastic IBM and a typically deterministic integrodifference model to estimate rates of spread (e.g., Travis et al., 2011; Santini et al., 2016). Notably, while until recently integrodifference models have almost exclusively been used to project spread rates across homogenous landscapes, recent developments are enabling rapid simulation of integrodifference equations across spatially complex landscapes (e.g., Synes et al., 2016; Gilbert et al., 2017). One major potential advantage of the integrodifference approach is that the much faster speed of individual simulations will make inverse fitting of parameters through Bayesian approaches including approximate Bayesian computation much more readily achievable. A further important development will be to integrate environmental niche modeling with the population dynamic modeling approaches available.

A key challenge is to move beyond the approach most often taken in what are often termed hybrid species distribution models and to relate the environmental variables directly to the key demographic traits (e.g., reproduction, survival, and dispersal), rather than simply using the environmental niche model to demark suitable and unsuitable environments for a focal species. However, many such relationships have yet to be established in detail (see Zurell et al., 2016 for excellent discussion of key issues). We note here that regardless of the modeling approach taken, in order to engage with stakeholders effectively, it is extremely useful to have clear spatially realistic model output that enables individuals of different backgrounds to relate to the modeling process and its potential (as called for in Stelzenmüller et al., 2018). Thus, as we develop more sophisticated and complex models for predicting and managing spread, we need also to focus on how we develop effective approaches for presenting the results of these models (including associated uncertainties) in an accessible form for those stakeholders with whom we are jointly developing the models, and for others who are likely to find the models useful.

CONCLUDING REMARK

We calibrated an IBM for the round goby, using spatial presence information from the invasion of the Gulf of Gdansk. Stakeholder involvement with question design provided both a preliminary answer and future research directions. It is important that we encourage a culture of publishing work on the process of co-development of models, such that we can learn from one another's successes and failures. This will require more papers, such as this one, that are published at potentially earlier stages of model development and before models are necessarily ready for use to inform management action. In this instance, while short of being ready to inform management action, the model has helped to emphasize the requirement for investment in gathering greater empirical understanding of the depth at which round goby disperse. In the next part of the co-development modeling spiral (Parrott, 2017) this will be gathered, and models will be built using this information, together with higher quality information on human-mediated dispersal pathways, to improve our ability to capture Baltic-wide patterns of invasion and to enable improved forecasts of future distribution under alternative management options to be developed. In general, we promote the increased use of models as a heuristic device for horizon scanning and risk assessment of invasive species and suggest that this utility may be at least as influential as their more traditional usage for informing management at the stage when they are well-validated.

AUTHOR CONTRIBUTIONS

Conceptualization: ES, JT, SP, TB, PH, and JB. Formal analysis: ES and SP. Investigation: ES, SP, JT, TB, PH, and JB. Methodology: ES and SP. Project administration: ES and JT. Resources: JT. Supervision: JT. Validation: ES and SP. Visualization: ES. Writing - original draft: ES, SP, JT, PH, JB. Writing - review and editing: ES, SP, PH, JB, JT, TB.

FUNDING

We are grateful to the Natural Environment Research Council (NE/J008001/1) for financial support. JB received funding from BONUS (BIO-C3 project), the joint Baltic Sea research and development programme (Art 185), funded jointly by the European Union's Seventh Framework Programme for Research, Technological Development and Demonstration and Innovation

Fund Denmark. In addition, PH and JB would like to thank the Swedish Agency for Marine and Water Management for covering travel expenses to the stakeholder workshop at Kalmar. This study is part of the project PROBIS (BiodivERsA2013) and financially supported by the Swedish Research Council Formas.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fevo.2017.00149/full#supplementary-material

REFERENCES

Aben, J., Bocedi, G., Palmer, S. C., Pellikka, P., Strubbe, D., Hallmann, C., et al. (2016). The importance of realistic dispersal models in conservation planning: application of a novel modelling platform to evaluate management scenarios in an Afrotropical biodiversity hotspot. *J. Appl. Ecol.* 53, 1055–1065. doi: 10.1111/1365-2664.12643

Adams, V. M., Petty, A. M., Douglas, M. M., Buckley, Y. M., Ferdinands, K. B., Okazaki, T., et al. (2015). Distribution, demography and dispersal model of spatial spread of invasive plant populations with limited data. *Methods Ecol. Evol.* 6, 782–794. doi: 10.1111/2041-210X.12392

Allouche, O., Tsoar, A., and Kadmon, R. (2006). Assessing the accuracy of species distribution models: prevalence, kappa and the true skill statistic (TSS). *J. Appl. Ecol.* 43, 1223–1232. doi: 10.1111/j.1365-2664.2006.01214.x

Almpanidou, V., Schofield, G., Kallimanis, A. S., Türkozan, O., Hays, G.C., and Mazaris, A.D. (2016). Using climatic suitability thresholds to identify past, present and future population viability. *Ecol. Indic.* 71, 551–556. doi: 10.1016/j.ecolind.2016.07.038

Azour, F., van Deurs, M., Behrens, J., Carl, H., Hüssy, K., Greisen, K., et al. (2015). Invasion rate and population characteristics of the round goby *Neogobius melanostomus*: effects of density and invasion history. *Aquat. Biol.* 24, 41–52. doi: 10.3354/ab00634

Bauduin, S., McIntire, E., St-Laurent, M.-H., and Cumming, S. (2016). Overcoming challenges of sparse telemetry data to estimate caribou movement. *Ecol. Model.* 335, 24–34. doi: 10.1016/j.ecolmodel.2016.05.004

Bayliss, H., Stewart, G., Wilcox, A., and Randall, N. (2013). A perceived gap between invasive species research and stakeholder priorities. *NeoBiota* 19, 67–82. doi: 10.3897/neobiota.19.4897

Behrens, J. W., van Deurs, M., and Christensen, E. A. F. (2017). Evaluating dispersal potential of an invasive fish by the use of aerobic scope and osmoregulation capacity. *PLoS ONE* 12:e0176038. doi: 10.1371/journal.pone.0176038

Bocedi, G., Palmer, S. C. F., Pe'er, G., Heikkinen, R. K., Matsinos, Y. G., Watts, K., et al. (2014). RangeShifter: a platform for modelling spatial eco-evolutionary dynamics and species' responses to environmental changes. *Methods Ecol. Evol.* 5, 388–396. doi: 10.1111/2041-210X.12162

Boje, J., Neuenfeldt, S., Sparrevohn, C. R., Eigaard, O., and Behrens, J. W. (2014). Seasonal migration, vertical activity, and winter temperature experience of Greenland halibut *Reinhardtius hippoglossoides* in West Greenland waters. *Mar. Ecol. Prog. Ser.* 508, 211–222. doi: 10.3354/meps10874

Brown, G. P., Phillips, B. L., and Shine, R. (2014). The straight and narrow path: the evolution of straight-line dispersal at a cane toad invasion front. *Proc. R. Soc. Lond. B Biol. Sci.* 281:20141385. doi: 10.1098/rspb.2014.1385

Brownscombe, J., Masson, L., Beresford, D., and Fox, M. (2012). Modeling round goby *Neogobius melanostomus* range expansion in a Canadian river system. *Aquat. Invasions* 7, 537–545. doi: 10.3391/ai.2012.7.4.010

Cardador, L., Carrete, M., Gallardo, B., and Tella, J. L. (2016). Combining trade data and niche modelling improves predictions of the origin and distribution of non-native European populations of a globally invasive species. *J. Biogeogr.* 43, 967–978. doi: 10.1111/jbi.12694

Carl, H., Behrens, J., and Møller, P. R. (2016). *Statusrapport Vedr. Udbredelsen af Ikke-Hjemmehørende Fiskearter i Danske Kystvande*. Statens Naturhistoriske Museum. Natural History Museum of Denmark, University of Copenhagen.

Charlebois, P. M., Corkum, L. D., Jude, D. J., and Knight, C. (2001). The round goby (*Neogobius melanostomus*) invasion: current research and future needs. *J. Great Lakes Res.* 27, 263–266. doi: 10.1016/S0380-1330(01)70641-7

Chotkowski, M. A., and Ellen Marsden, J. (1999). Round goby and mottled sculpin predation on lake trout eggs and fry: field predictions from laboratory experiments. *J. Great Lakes Res.* 25, 26–35. doi: 10.1016/S0380-1330(99)70714-8

Crowl, T. A., Crist, T. O., Parmenter, R. R., Belovsky, G., and Lugo, A. E. (2008). The spread of invasive species and infectious disease as drivers of ecosystem change. *Front. Ecol. Environ.* 6:70151. doi: 10.1890/070151

Cuddington, K., Fortin, M. J., Gerber, L. R., Hastings, A., Liebhold, A., O'Connor, M., et al. (2013). Process-based models are required to manage ecological systems in a changing world. *Ecosphere* 4, 1–12. doi: 10.1890/ES12-00178.1

Ellis, S., and Macisaac, H. J. (2008). Salinity tolerance of Great Lakes invaders. *Freshwater Biol.* 54, 77–89. doi: 10.1111/j.1365-2427.2008.02098.x

Evans, M. R., Bithell, M., Cornell, S. J., Dall, S. R. X., Díaz, S., Emmott, S., et al. (2013). Predictive systems ecology. *Proc. R. Soc. Lond. B Biol. Sci.* 280:20131452. doi: 10.1098/rspb.2013.1452

Franklin, J. (2010). Moving beyond static species distribution models in support of conservation biogeography. *Divers. Distrib.* 16, 321–330. doi: 10.1111/j.1472-4642.2010.00641.x

Fraser, E. J., Lambin, X., Travis, J. M. J., Harrington, L. A., Palmer, S. C. F., Bocedi, G., et al. (2015). Range expansion of an invasive species through a heterogeneous landscape – the case of American mink in Scotland. *Divers. Distrib*, 21, 888–900. doi: 10.1111/ddi.12303

Freeman, E., and Moisen, G. (2008). presenceabsence: an r package for presence-absence model analysis. *J. Stat. Softw.* 23:31. doi: 10.18637/jss.v023.i11

Gallardo, B., Clavero, M., Sánchez, M. I., and Vilà, M. (2016). Global ecological impacts of invasive species in aquatic ecosystems. *Glob. Change Biol.* 22, 151–163. doi: 10.1111/gcb.13004

Gilbert, M. A., White, S. M., Bullock, J. M., and Gaffney, E. A. (2017), Speeding up the simulation of population spread models. *Methods Ecol. Evol.* 8, 501–510. doi: 10.1111/2041-210X.12684

Goldstein, E. A., Butler, F., and Lawton, C. (2016). Modeling future range expansion and management strategies for an invasive squirrel species. *Biol. Invasions* 18, 1431–1450. doi: 10.1007/s10530-016-1092-7

Grimm, V., and Railsback, S. F. (2012). Pattern-oriented modelling: a multi-scope for predictive systems ecology. *Philos. Trans. R. Soc. Lond. B Biol. Sci.* 367, 298–310. doi: 10.1098/rstb.2011.0180

Grimm, V., Revilla, E., Berger, U., Jeltsch, F., Mooij, W. M., Railsback, S. F., et al. (2005). Pattern-oriented modeling of agent-based complex systems: lessons from ecology. *Science* 310, 987–991. doi: 10.1126/science.1116681

Groen, M., Sopinka, N. M., Marentette, J. R., Reddon, A. R., Brownscombe, J. W., Fox, M. G., et al. (2012). Is there a role for aggression in round goby invasion fronts? *Behaviour* 149, 685–703. doi: 10.1163/1568539X-00002998

Grygiel, W. (2007). *Round Goby (Neogobius melanostomus Pallas, 1811) "Semi-domestic" Species in the Gulf of Gdansk the Southern Baltic; 1993-2004).* ICES CM 2007/E:30; Marine biodiversity: A fish and fisheries perspective.

Hadorn, G. H., Hoffmann-Riem, H., Biber-Klemm, S., Grossenbacher-Mansuy, W., Joye, D., Pohl, C., et al. (2008). *Handbook of Transdisciplinary Research*. Springer Netherlands. doi: 10.1007/978-1-4020-6699-3

Hastings, A., Cuddington, K., Davies, K. F., Dugaw, C. J., Elmendorf, S., Freestone, A., et al. (2005). The spatial spread of invasions: new developments in theory and evidence. *?Ecol. Lett.* 8, 91–101. doi: 10.1111/j.1461-0248.2004.00687.x

Hearn, R. D., Harrison, A. L., and Cranswick, P. A. (2015). *Draft International Single Species Action Plan for the Conservation of the Long-Tailed Duck Clangula hyemalis*, 2016–2025. AEWA Technical Series.

Heikkinen, R. K., Bocedi, G., Kuussaari, M., Heliölä, J., Leikola, N., Pöyry, J., et al. (2014). Impacts of land cover data selection and trait parameterisation on dynamic modelling of species' range expansion. *PLoS ONE* 9:e108436. doi: 10.1371/journal.pone.0108436

Higgins, S. I., Richardson, D. M., and Cowling, R. M. (1996). Modeling invasive plant spread: the role of plant-environment interactions and model structure. *Ecology* 77, 2043–2054. doi: 10.2307/2265699

Hirsch, P. E., Adrian-Kalchhauser, I., Flämig, S., N'Guyen, A., Defila, R., Di Giulio, A., et al. (2016a). A tough egg to crack: recreational boats as vectors for invasive goby eggs and transdisciplinary management approaches. *Ecol. Evol.* 6, 707–715. doi: 10.1002/ece3.1892

Hirsch, P. E., N'Guyen, A., Adrian-Kalchhauser, I., and Burkhardt-Holm, P. (2016c). What do we really know about the impacts of one of the 100 worst invaders in Europe? A reality check. *Ambio* 45, 267–279. doi: 10.1007/s13280-015-0718-9

Hirsch, P. E., Thorlacius, M., Brodin, T., and Burkhardt-Holm, P. (2017). An approach to incorporate individual personality in modeling fish dispersal across in-stream barriers. *Ecol. Evol.* 7, 720–732. doi: 10.1002/ece3.2629

Hirsch, P., Burkhardt-Holm, P., Töpfer, I., and Fischer, P. (2016b). Movement patterns and shelter choice of spiny-cheek crayfish (*Orconectes limosus*) in a large lake's littoral zone. *Aquat. Invasions* 11, 55–65. doi: 10.3391/ai.2016.11.1.06

Huang, F., Peng, S., Chen, B., Liao, H., Huang, Q., Lin, Z., et al. (2015). Rapid evolution of dispersal-related traits during range expansion of an invasive vine *Mikania micrantha*. *Oikos* 124, 1023–1030. doi: 10.1111/oik.01820

Hudson, C. M., Phillips, B. L., Brown, G. P., and Shine, R. (2015). Virgins in the vanguard: low reproductive frequency in invasion-front cane toads. *Biol. J. Linnean Soc.* 116, 743–747. doi: 10.1111/bij.12618

Hulme, P. E. (2009). Trade, transport and trouble: managing invasive species pathways in an era of globalization. *J. Appl. Ecol.* 46, 10–18. doi: 10.1111/j.1365-2664.2008.01600.x

Huntley, B., Barnard, P., Altwegg, R., Chambers, L., Coetzee, B. W. T., Gibson, L., et al. (2010). Beyond bioclimatic envelopes: dynamic species' range and abundance modelling in the context of climatic change. *Ecography* 33, 621–626. doi: 10.1111/j.1600-0587.2009.06023.x

Karlson, A. M. L., Almqvist, G., Skóra, K. E., and Appelberg, M. (2007). Indications of competition between non-indigenous round goby and native flounder in the Baltic Sea. *ICES J. Mar. Sci.* 64, 479–486. doi: 10.1093/icesjms/fsl049

Karsiotis, S. I., Pierce, L. R., Brown, J. E., and Stepien, C. A. (2012). Salinity tolerance of the invasive round goby: experimental implications for seawater ballast exchange and spread to North American estuaries. *J. Great Lakes Res.* 38, 121–128. doi: 10.1016/j.jglr.2011.12.010

Katsanevakis, S., Mackelworth, P., Coll, M., Fraschetti, S., Mačić, V., Giakoumi, S., et al. (2017). Advancing marine conservation in European and contiguous seas with the MarCons Action. *Res. Ideas Outcomes* 3:e11884. doi: 10.3897/rio.3.e11884

Kornis, M. S., Mercado-Silva, N., and Vander Zanden, M. J. (2012). Twenty years of invasion: a review of round goby *Neogobius melanostomus* biology, spread and ecological implications. *J. Fish Biol.* 80, 235–285. doi: 10.1111/j.1095-8649.2011.03157.x

Kotta, J., Nurkse, K., Puntila, R., and Ojaveer, H. (2016). Shipping and natural environmental conditions determine the distribution of the invasive non-indigenous round goby *Neogobius melanostomus* in a regional sea. *Estuar. Coast. Shelf Sci.* 169, 15–24. doi: 10.1016/j.ecss.2015.11.029

Krueger, T., Page, T., Hubacek, K., Smith, L., and Hiscock, K. (2012). The role of expert opinion in environmental modelling. *Environ. Model. Softw.* 36, 4–18. doi: 10.1016/j.envsoft.2012.01.011

Kwon, T.-D., Fisher, S. W., Kim, G. W., Hwang, H., and Kim, J.-E. (2006). Trophic transfer and biotransformation of polychlorinated biphenyls in zebra mussel, round goby, and smallmouth bass in Lake Erie, U.S.A. *Environ. Toxicol. Chem.* 25, 1068–1078. doi: 10.1897/05-180R.1

Ladago, B. J., Marsden, J. E., and Evans, A. N. (2016). Early feeding by lake trout fry. *Trans. Am. Fish. Soc Trans. Am. Fish. Soc.* 145, 1–6. doi: 10.1080/00028487.2015.1073622

Lurgi, M., Brook, B. W., Saltré, F., and Fordham, D. A. (2015). Modelling range dynamics under global change: which framework and why? *Methods Ecol. Evol.* 6, 247–256. doi: 10.1111/2041-210X.12315

Marentette, J. R., Wang, G., Tong, S., Sopinka, N. M., Taves, M. D., Koops, M. A., et al. (2011). Laboratory and field evidence of sex-biased movement in the invasive round goby. *Behav. Ecol. Sociobiol.* 65, 2239–2249. doi: 10.1007/s00265-011-1233-z

Matzek, V., Pujalet, M., and Cresci, S. (2015). What managers want from invasive species research versus what they get. *Conserv. Lett.* 8, 33–40. doi: 10.1111/conl.12119

Meekins, J. F., and Mccarthy, B. C. (2002). Effect of population density on the demography of an invasive plant (*Alliaria petiolata*, brassicaceae) population in a Southeastern Ohio Forest. *Am. Midl. Nat.* 147, 256–278. doi: 10.1674/0003-0031(2002)147[0256:EOPDOT]2.0.CO;2

Michalek, M., Puntila, R., Strake, S., and Werner, M. (2012). *Abundance and Distribution of Round Goby* (*Neogobius melanostomus*). Available online at: http://www.helcom.fi/baltic-sea-trends/environment-fact-sheets (Accessed September 20, 2016).

Molnar, J. L., Gamboa, R. L., Revenga, C., and Spalding, M. D. (2008). Assessing the global threat of invasive species to marine biodiversity. *Front. Ecol. Environ.* 6:70064. doi: 10.1890/070064

Myles-Gonzalez, E., Burness, G., Yavno, S., Rooke, A., and Fox, M. G. (2015). To boldly go where no goby has gone before: boldness, dispersal tendency, and metabolism at the invasion front. *Behav. Ecol.* 26, 1083–1090. doi: 10.1093/beheco/arv050

N'Guyen, A., Hirsch, P. E., Adrian-Kalchhauser, I., and Burkhardt-Holm, P. (2016). Improving invasive species management by integrating priorities and contributions of scientists and decision makers. *Ambio* 45, 280–289. doi: 10.1007/s13280-015-0723-z

N'Guyen, A., Hirsch, P. E., Bozzuto, C., Adrian-Kalchhauser, I., Hôrková, K., and Burkhardt-Holm, P. (2017). A dynamical model for invasive round goby populations reveals efficient and effective management options. *J. Appl. Ecol.* doi: 10.1111/1365-2664.12934. [Epub ahead of print].

Ojaveer, H., Galil, B., Lehtiniemi, M., Christoffersen, M., Clink, S., Florin, A.-B., et al. (2015). Twenty five years of invasion: management of the round goby *Neogobius melanostomus* in the Baltic Sea. *Manag. Biol. Invasion* 6, 329–339. doi: 10.3391/mbi.2015.6.4.02

Palmer, S. C. F., Coulon, A., and Travis, J. M. J. (2011). Introducing a 'stochastic movement simulator' for estimating habitat connectivity. *Methods Ecol. Evol.* 2, 258–268. doi: 10.1111/j.2041-210X.2010.00073.x

Parrott, L. (2017). The modelling spiral for solving "wicked" environmental problems: guidance for stakeholder involvement and collaborative model development. *Methods Ecol. Evol.* 8, 1005–1011. doi: 10.1111/2041-210X.12757

Parrott, L., Chion, C., Gonzalès, R., and Latombe, G. (2012). Agents, individuals, and networks: modeling methods to inform natural resource management in regional landscapes. *Ecol. Soc.* 17, 32. doi: 10.5751/ES-04936-170332

Parry, H. R., Eagles, D., and Kriticos, D. J. (2015). "Simulation modelling of long-distance windborne dispersal for invasion ecology," in *Pest Risk Modelling and Mapping for Invasive Alien Species,* ed R. C. Venette (Wallingford: CABI). doi: 10.1079/9781780643946.0049

Phillips, B. L., Brown, G. P., Webb, J. K., and Shine, R. (2006). Invasion and the evolution of speed in toads. *Nature* 439, 803–803. doi: 10.1038/439803a

Pimentel, D., Lach, L., Zuniga, R., and Morrison, D. (2000). Environmental and economic costs of nonindigenous species in the United States. *BioScience* 50, 53–65. doi: 10.1641/0006-3568(2000)050[0053:EAECON]2.3.CO;2

Pimentel, D., Zuniga, R., and Morrison, D. (2005). Update on the environmental and economic costs associated with alien-invasive species in the United States. *Ecol. Econ.* 52, 273–288. doi: 10.1016/j.ecolecon.2004.10.002

Pohl, C., and Hadorn, G. H. (2008). Methodological challenges of transdisciplinary research. *Nat. Sci. Soc.* 16, 111–121. doi: 10.1051/nss:2008035

Reed, M., Hubacek, K., Bonn, A., Burt, T., Holden, J., Stringer, L., et al. (2013). Anticipating and managing future trade-offs and complementarities between ecosystem services. *Ecol. Soc.* 18, 5. . doi: 10.5751/ES-04924-180105

Salo, P., Korpimäki, E., Banks, P. B., Nordström, M., and Dickman, C. R. (2007). Alien predators are more dangerous than native predators to prey populations. *Proc. R. Soc. B* 274, 1237–1243. doi: 10.1098/rspb.2006.0444

Santini, L., Cornulier, T., Bullock, J. M., Palmer, S. C., White, S. M., Hodgson, J. A., et al. (2016). A trait-based approach for predicting species responses to environmental change from sparse data: how well might terrestrial mammals track climate change? *Glob. Change Biol.* 22, 2415–2424. doi: 10.1111/gcb.13271

Sapota, M. R. (2004). The round goby (*Neogobius melanostomus*) in the Gulf of Gdansk – a species introduction into the Baltic Sea. *Hydrobiologia* 514, 219–224. doi: 10.1023/B:hydr.0000018221.28439.ae

Sapota, M. R. (2012). *NOBANIS – Invasive Alien Species Fact Sheet – Neogobius melanostomus. From: Online Database of the European Network on Invasive Alien Species–NOBANIS.* Available online at: www.nobanis.org (Accessed: July 20 2017).

Schrandt, M. N., Stone, L. C., Klimek, B., Mäkelin, S., Heck Jr, K. L., Mattila, J., et al. (2016). A laboratory study of potential effects of the invasive round goby on nearshore fauna of the Baltic Sea. *Aquat. Invasions* 11, 327–335. doi: 10.3391/ai.2016.11.3.10

Seidl, R., Brand, F. S., Stauffacher, M., Krütli, P., Le, Q. B., Spörri, A., et al. (2013). Science with society in the Anthropocene. *AMBIO* 42, 5–12. doi: 10.1007/s13280-012-0363-5

Skabeikis, A., and Lesutiene, J. (2015). Feeding activity and diet composition of round goby (*Neogobius melanostomus*, Pallas 1814) in the coastal waters of SE Baltic Sea. *Oceanol. Hydrobiol. Stud*. 44, 508–519. doi: 10.1515/ohs-2015-0048

Smith, J. M., and Slatkin, M. (1973). The stability of predator-prey systems. *Ecology* 54, 384–391. doi: 10.2307/1934346

Stelzenmüller, V., Coll, M., Mazaris, A. D., Giakoumi, S., Katsanevakis, S., Portman, M. E., et al. (2018). A risk-based approach to cumulative effect assessments for marine management. *Sci. Tot. Environ*. 612, 1132–1140. doi: 10.1016/j.scitotenv.2017.08.289

Synes, N. W., Brown, C., Watts, K., White, S. M., Gilbert, M. A., and Travis, J. M. J. (2016). Emerging opportunities for landscape ecological modelling. *Curr. Landscape Ecol. Rep*. 1, 146–167. doi: 10.1007/s40823-016-0016-7

Therry, L., Bonte, D., and Stoks, R. (2015). Higher investment in flight morphology does not trade off with fecundity estimates in a poleward range-expanding damselfly. *Ecol. Entomol*. 40, 133–142. doi: 10.1111/een.12170

Thorlacius, M., Hellström, G., and Brodin, T. (2015). Behavioral dependent dispersal in the invasive round goby *Neogobius melanostomus* depends on population age. *Curr. Zool*. 61, 529–542. doi: 10.1093/czoolo/61.3.529

Thuiller, W., Münkemüller, T., Lavergne, S., Mouillot, D., Mouquet, N., Schiffers, K., et al. (2013). A road map for integrating eco-evolutionary processes into biodiversity models. *?Ecol. Lett*. 16, 94–105. doi: 10.1111/ele.12104

Tierney, K., Kasurak, A., Zielinski, B., and Higgs, D. (2011). Swimming performance and invasion potential of the round goby. *Environ. Biol. Fish*. 92, 491–502. doi: 10.1007/s10641-011-9867-2

Travis, J. M. J., Harris, C. M., Park, K. J., and Bullock, J. M. (2011). Improving prediction and management of range expansions by combining analytical and individual-based modelling approaches. *Methods Ecol. Evol*. 2, 477–488. doi: 10.1111/j.2041-210X.2011.00104.x

Travis, J. M., Münkemüller, T., and Burton, O. J. (2010). Mutation surfing and the evolution of dispersal during range expansions. *J. Evol. Biol*. 23, 2656–2667. doi: 10.1111/j.1420-9101.2010.02123.x

Uden, D. R., Allen, C. R., Angeler, D. G., Corral, L., and Fricke, K. A. (2015). Adaptive invasive species distribution models: a framework for modeling incipient invasions. *Biol. Invasions* 17, 2831–2850. doi: 10.1007/s10530-015-0914-3

Urban, M. C., Bocedi, G., Hendry, A. P., Mihoub, J.B., Pe'er, G., Singer, A., et al. (2016). Improving the forecast for biodiversity under climate change. *Science* 353:aad8466. doi: 10.1126/science.aad8466

Vélez-Espino, L. A., Koops, M. A., and Balshine, S. (2010). Invasion dynamics of round goby (*Neogobius melanostomus*) in Hamilton Harbour, Lake Ontario. *Biol. Invasions* 12, 3861–3875. doi: 10.1007/s10530-010-9777-9

Voinov, A., and Bousquet, F. (2010). Modelling with stakeholders. *Environ. Model. Softw*. 25, 1268–1281. doi: 10.1016/j.envsoft.2010.03.007

Vuilleumier, S., Buttler, A., Perrin, N., and Yearsley, J. M. (2011). Invasion and eradication of a competitively superior species in heterogeneous landscapes. *Ecol. Model*. 222, 398–406. doi: 10.1016/j.ecolmodel.2010.09.037

Wittmann, M. E., Chandra, S., Reuter, J. E., Schladow, S. G., Allen, B. C., and Webb, K. J. (2012). The control of an invasive bivalve, *Corbicula fluminea*, using gas impermeable benthic barriers in a large natural lake. *Environ. Manage*. 49, 1163–1173. doi: 10.1007/s00267-012-9850-5

Zurell, D., Thuiller, W., Pagel, J., Cabral, J. S., Münkemüller, T., Gravel, D., et al. (2016). Benchmarking novel approaches for modelling species range dynamics. *Glob. Change Biol*. 22, 2651–2664. doi: 10.1111/gcb.13251

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers
in Ecology and Evolution

PERSPECTIVE
published: 07 December 2017
doi: 10.3389/fevo.2017.00158

Using Environmental DNA to Improve Species Distribution Models for Freshwater Invaders

Teja P. Muha[1]*, *Marta Rodríguez-Rey*[1]*, *Matteo Rolla*[1] *and Elena Tricarico*[2]

[1] *Bioscience Department, Swansea University, Swansea, United Kingdom,* [2] *Department of Biology, University of Florence, Florence, Italy*

Species Distribution Models (SDMs) have been reported as a useful tool for the risk assessment and modeling of the pathways of dispersal of freshwater invasive alien species (IAS). Environmental DNA (eDNA) is a novel tool that can help detect IAS at their early stage of introduction and additionally improve the data available for a more efficient management. SDMs rely on presence and absence of the species in the study area to infer the predictors affecting species distributions. Presence is verified once a species is detected, but confirmation of absence can be problematic because this depends both on the detectability of the species and the sampling strategy. eDNA is a technique that presents higher detectability and accuracy in comparison to conventional sampling techniques, and can effectively differentiate between presence or absence of specific species or entire communities by using a barcoding or metabarcoding approach. However, a number of potential bias can be introduced during (i) sampling, (ii) amplification, (iii) sequencing, or (iv) through the usage of bioinformatics pipelines. Therefore, it is important to report and conduct the field and laboratory procedures in a consistent way, by (i) introducing eDNA independent observations, (ii) amplifying and sequencing control samples, (iii) achieving quality sequence reads by appropriate clean-up steps, (iv) controlling primer amplification preferences, (v) introducing PCR-free sequence capturing, (vi) estimating primer detection capabilities through controlled experiments and/or (vii) *post-hoc* introduction of "site occupancy-detection models." With eDNA methodology becoming increasingly routine, its use is strongly recommended to retrieve species distributional data for SDMs.

Keywords: aquatic freshwater invasive species, barcoding, metabarcoding, environmental DNA, environmental sampling, independent evaluation

Edited by:
Stelios Katsanevakis,
University of the Aegean, Greece

Reviewed by:
Emre Keskin,
Ankara University, Turkey
Mikkel Winther Pedersen,
University of Cambridge,
United Kingdom

Correspondence:
Teja P. Muha
t.p.muha@swansea.ac.uk
Marta Rodríguez-Rey
m.rodriguez-reygomez@swansea.ac.uk

Specialty section:
This article was submitted to
Environmental Informatics,
a section of the journal
Frontiers in Ecology and Evolution

Received: *29 September 2017*
Accepted: *24 November 2017*
Published: *07 December 2017*

Citation:
Muha TP, Rodríguez-Rey M, Rolla M and Tricarico E (2017) Using Environmental DNA to Improve Species Distribution Models for Freshwater Invaders. Front. Ecol. Evol. 5:158. doi: 10.3389/fevo.2017.00158

INTRODUCTION

Current policies on invasive alien species (IAS) depend on the availability and quality of data used for their risk assessment (Groom et al., 2017). Species Distribution Models (SDMs) use available data of invasive species and are one of the most widely used tools for risk assessment, predicting species distribution and pathways of dispersal (Jiménez-Valverde et al., 2011).

This methodology relates the distribution data of the IAS (e.g., presence and absence records) in the study area with a set of independent spatially explicit variables to explain and predict the range expansion of the species. However, there are limitations on these approaches because of two main reasons: (i) confirmed absences are desirable but scarce in available databases, and (ii) independent data for evaluation is normally not available. The consideration of absences has been reported to provide more accurate predictions of the actual distribution of IAS (Václavík and Meentemeyer, 2009). Therefore, there is a need for tools that allow the recording of presence and absence and a faster compilation of independent data to test spatially explicit models. Efficient spatial monitoring of invasive species vectors of introduction, further dispersal as well as initial detection of newly present species, are crucial for species management as are prevention, control and eradication.

In the recent years, a new environmental molecular tool has been developed- environmental DNA (eDNA). eDNA refers to DNA which can be extracted from environmental samples without separation of specific organisms from the environment (Taberlet et al., 2012). eDNA contains both cellular as well as extracellular DNA from all kinds of organisms. It is subject to high levels of degradation but can be preserved in nature from few weeks up to hundreds of thousands of years (Thomsen and Willerslev, 2015). The ability to detect species through eDNA water samples is relatively novel and has proved as a useful tool for the detection of aquatic IAS (Dejean et al., 2012; Goldberg et al., 2013; Nathan et al., 2014). It can be applied for the detection of a number of specific IAS (barcoding), or detecting multiple IAS as part of whole communities (metabarcoding). New revolutionary techniques for eDNA are being developed on a daily basis with the aim to provide a number of useful information such as, presence or absence of the species (Ficetola et al., 2008), density assessments (Moyer et al., 2014), population dynamics (Sigsgaard et al., 2016), sex (Nichols and Spong, 2017), hybridization process between subspecies, (Uchii et al., 2016; Goricki et al., 2017), spatial representativeness (Civade et al., 2016; Bista et al., 2017) and ability to amplify whole mitochondrial genome (Deiner et al., 2017b). A wide range of eDNA detection possibilities is currently limited. Knowing what are the limitations of eDNA methods is key to successful estimation of species presence (or absence) and estimations of their biological characteristics.

APPROACH

Nowadays, useful information on IAS within SDMs is in the detection of presence and absence of the species (Ficetola et al., 2008). In this article, we discuss the range of possibilities and limitations with regard to reporting IAS presence or absence using eDNA in freshwater ecosystems in order to obtain additional and more accurate distribution data to be used in the SDMs.

Potential Applications

eDNA has thus far been mainly used in the early detection and monitoring of invasive species, contributing to the increase of IAS presence records. The use of eDNA techniques could facilitate a more effective method for recording IAS absence than do regular monitoring surveys or possibly may aid in the compilation of independent data similar to the approach used for proving (non)successful eradications (Dejean et al., 2012). Currently, eDNA research is focusing its effort on the species detection efficiencies based on the competence of sampling, amplification and sequencing techniques. A detailed review has been conducted based on the potential for the future application of eDNA tool by identifying the proportion of positive detections of IAS within individual research (**Table 1**). The review proves how useful the tool can be dealing with IAS detection. A recent increase in presented eDNA research conducted on invasive species is only the tip of the iceberg of what can be achieved for conservation and IAS management. There is however a number of limitations that should be remembered before applying eDNA data to retrieve distribution data for SDMs.

Current Limitations

Freshwater ecosystems, lentic, and lotic, provide excellent study area for defining the wide range of detection possibilities of eDNA techniques as well as the limitations. Small-scale freshwater lentic bodies provide an excellent opportunity to study eDNA characteristics related to degradation, which can affect successful detectability of species. Recent studies have tried to underline degradation rates in correlation to abiotic factors, such as, (i) most effective water stratum for eDNA detection (Moyer et al., 2014), (ii) pH, UV-B (Strickler et al., 2015), (iii) effects of temperature on eDNA degradation (Strickler et al., 2015; Eichmiller et al., 2016), and (iv) temporal effects (Dejean et al., 2011). Freshwater lotic bodies can provide important information due to their longitudinal downstream dynamics, such as, (i) eDNA persistence in the environment (Jerde et al., 2016; Wilcox et al., 2016), (ii) residence time of eDNA (Jerde et al., 2016), and (iii) the ecology of eDNA (Barnes and Turner, 2016). In case of newly introduced IAS, measures of low abundances present another limitation (Jerde et al., 2011) which is highly important when discerning between presence and absence records. Some of the reported examples are applied to non-invasive species, but the reason why we focus on IAS is that time, i.e., rapid response, is key to management, so that an identified IAS can be eradicated/controlled before any negative ecosystem impact occurs. Since eDNA can assist in more rapid detection and early response to IAS invasions than traditional sampling, this technology most greatly benefits identification of invasive species.

All the limitations of eDNA that are currently being studied are crucial for IAS assessment. When monitoring, especially in a new environment, it is fundamental to detect it at extremely low abundances and report negative or positive presence. False positives and negatives are essentially relevant for their use within SDM and cannot be misjudged, whether they are products of sampling bias or metabarcoding bioinformatics pipeline. The distribution patterns and biology of the eDNA is another important factor influencing the accuracy of information which is relevant for the distribution of IAS within the models. The accuracy that we can obtain through eDNA highly depends on the strategies followed during the fieldwork and through laboratory protocols. In order to more accurately state the proportion of the positive (or negative) detections, independent observations (Steel et al., 2013) would need to become an essential part of eDNA studies to overcome the bias of false

TABLE 1 | eDNA studies targeting freshwater invasive alien species, including description of water sampling and filtration techniques, DNA loci, barcoding or metabarcoding as well as the proportion of positive detections.

Taxon	Target freshwater IAS	Sampling technique; filtration or precipitation procedure	DNA loci	eDNA amplification/sequencing method	Proportion of positive detections (%)	References
Insects	Tiger mosquito, *Aedes albopictus* Asian bush mosquito, *Aedes japonicus japonicas; Aedes koreicus*	Collection of 3 × 15 ml; Ethanol precipitation (EP) (15 mL of water + 1.5 mL of sodium acetate 3 M and 33 ml absolute ethanol) Precipitation of DNA by centrifuge (5,500 g, 35 min, 6°C) Ficetola et al., 2008	Ribosomal internal transcribed spacer 1 (ITS 1) Cytochrome oxidase subunit I (COI)	Quantitative real-time PCR (qPCR) + DNA metabarcoding	100% cPCR; 80% DNA metabarcoding	Schneider et al., 2016
Macrophytes	Brazilian waterweed, *Egeria densa*	EP –centrifuge by (20 min at 5,350 g) Ficetola et al., 2008	trnL– trnF	qPCR	Detected in all the ponds where it was observed.	Fujiwara et al., 2016
Reptiles	Burmese python, *Python bivittatus*	EP –centrifuge by (20 min at 5,350 g) Ficetola et al., 2008	Cyt b gene	Conventional PCR (cPCR)	100% (detected in the 5 sites where it has been observed)	Piaggio et al., 2014
Amphibians	American bullfrog, *Lithobates catesbeianus*	EP- Ficetola et al., 2008	Cyt b gene	cPCR	77.5% by eDNA, 14.3% by traditional methods (eDNA method indicated bullfrog occurrence in 38 out of 49 ponds.	Dejean et al., 2012
	Chinese giant salamander, *Andrias davidianus*	One 4-L container of surface water sample was collected per site; Glass fiber filter (0.7 μm)	mt NADH-1	Real-time TaqMan® PCR	Detected in 9 over 37 sites.	
	African clawed frog, *Xenopus laevis*	20 water samples of 40 ml per site; EP by Ficetola et al., 2008	12s *rRNA*	qPCR	Mean: 83%	Secondi et al., 2016
	American bullfrog, *Lithobates catesbeianus*	One 250 mL water sample per tank; polycarbonate filters (1.2 μm)	12s *rRNA*	DNA metabarcoding	10/12 tanks	Dejean et al., 2012
Crustaceans	Red swamp crayfish, *Procambarus clarkii*	Twenty 40 ml water samples per pond; EP Ficetola et al., 2008	COI	qPCR	eDNA 73%, trapping 65%	Tréguier et al., 2014
	Signal crayfish, *Pacifastacus leniusculus* Rusty crayfish, *Orconectes rusticus*	Five to ten water samples of 250 ml per site; cellulose nitrate filters (1.2 μm)	COI	qPCR	Weak relationships between eDNA copy number for *P. leniusculus* and relative abundance as catch per unit effort (CPUE)	Larson et al., 2017
	Rusty crayfish, *Orconectes rusticus*	Ten 250 mL surface water samples per site; cellulose nitrate or polycarbonate track-etch filters (1.2 μm)	COI	qPCR	Detection probability 95% at moderate-high abundance	Dougherty et al., 2016
Mollusc	New Zealand mudsnails, *Potamopyrgus antipodarum*	Three 4 L water samples per site; mixed cellulose ester membranes (0.45 μm)	COI	qPCR	Species detected in all 3 water samples from the first site and in 2 of 3 in the second site.	Goldberg et al., 2013

(Continued)

TABLE 1 | Continued

Taxon	Target freshwater IAS	Sampling technique; filtration or precipitation procedure	DNA loci	eDNA amplification/sequencing method	Proportion of positive detections (%)	References
Fish	Bluegill sunfish, *Lepomis macrochirus*	1 L water sample from the surface of each pond; cellulose acetate filter (3.0 μm)	COI	qPCR	Species found in 19 over 70 ponds, with traditional methods only 8 over 70 ponds.	Takahara et al., 2013
	Common carp, *Cyprinus carpio* Redfin perch, *Perca fluviatilis* Oriental weatherloach, *Misgurnus anguillicaudatus*	Six 2 L water samples per site; glass fiber filters (1.2 μm)	12S rRNA	qPCR	No significant correlation between catch per unit effort (CPUE) and DNA Positive correlation between CPUE and DNA Positive correlation between CPUE and DNA	Hinlo et al., 2017
	Common carp, *Cyprinus carpio* Eastern mosquitofish, *Gambusia holbrooki*	One 250 mL water sample per tank; polycarbonate membrane filters (1.2 μm)	12s + 16s rRNA	DNA metabarcoding	NA 3/12 tanks	Evans et al., 2016
	Common carp, *Cyprinus carpio*	One 50 mL water sample per tank; polycarbonate filter (0.2 μm)	COI	Multiplex qPCR	NA	Eichmiller et al., 2016
	Silver carp, *Hypophthalmichthys molitrix* Bighead carp, *Hypophthalmichthys nobilis*	2 L water sample; glass fiber filter (1.5 μm)	mtDNA D-loop	cPCR	Consistent with the traditional surveys	Jerde et al., 2013
	Common carp, *Cyprinus carpio* Rainbow trout, *Oncorhynchus mykiss* Minnow, *Phoxinus phoxinus* Brown trout, *Salmo trutta* Pike, *Esox Lucius*	36 × 2 L samples in three lakes; cellulose nitrate filter (0.45 μm)	CytB + 12S	eDNA metabarcoding	NA	Hänfling et al., 2016
	Common carp, *Cyprinus carpio*	One 500 mL water sample per tank; glass fiber filter (0.7 μm)	mtDNA D-loop	qPCR	NA	Uchii et al., 2016
	Pike, *Esox lucius*	Ten 1 L water samples; nitrocellulose mixed ester membrane (0.45–1.5 μm)	COI	qPCR	90% success rate	Dunker et al., 2016
	Northern snakehead, *Channa argus*	211 water samples in 7 locations; glass microfiber filters (1.5 μm)	16S	ddPCR	NA	Simmons et al., 2015
	Ruffe, *Gymnocephalus cernua*	2-L water samples from 24 locations; glass microfiber filters (1.5 μm)	COI	qPCR	Consistently higher success rate compared to conventional sampling	Tucker et al., 2016
	Round Goby, *Neogobius melanostomus*	500 mL water samples; glass microfiber filters (1.2 μm)	COI	eDNA metabarcoding	Out of 82 fish species—eDNA methods detected 86.2 and 72.0% in two rivers.	Balasingham et al., 2017

positives or negatives. An increased eDNA sampling effort based on a temporary scale would provide a more accurate proportion of positive (negative) detections and should be replaced by research proposed on a single sampling events (Simmons et al., 2015; Fujiwara et al., 2016; Hänfling et al., 2016). Independent observations would need to become a necessary procedure especially when dealing with estimations of newly introduced species (Jerde et al., 2011) or dealing with the estimations of successful eradication measures (Dunker et al., 2016).

To avoid bias due to inconsistent use of eDNA tools a minimum information based on field and laboratory procedures should always be reported and presented in a consistent manner as presented by (Goldberg et al., 2016). Pioneers in eDNA research (Ficetola et al., 2016) highly recommend following general requirements such as, precautionary approach to avoid contamination, respecting a general practice of obtaining control samples, extraction blanks, as well as incorporating PCR positive and negative controls. In cases of individual species assessment, parallel mesocom experiments are highly recommended in order to be able to estimate the limitations of detectability for each individual primer set. Another method to assess limitations of primer detections is assessing detectability of the species "in time" after its removal from the controlled environment. When working on multiple species assessment using a metabarcoding approach, it is recommended, to sequence the control samples, compare the sequencing control outputs with the actual samples, and if none of the last achieve high quality sequence reads by appropriate clean up steps; removal of singletons, chimeras, as well as including a record of removed sequences (Deiner et al., 2017a). Bias due to universal primer preferential amplifications of species can alter the relative abundance of individual species eDNA (Deiner et al., 2017a). A PCR-free method, namely sequence capturing offers promising solutions in order to avoid amplification bias (Shokralla et al., 2016).

In terms of IAS certainty of existence in a non-native environment, false- positive and false- negative are crucial points for management and environmental policies (Moyer et al., 2014; Lahoz-Monfort et al., 2016). Even low rate false- positives pose a bias toward species specific occupancy (Lahoz-Monfort et al., 2016). Errors produced during PCR and sequencing are main source of bias for false- positives whereas false- negatives normally appear due to bias during sampling. Sampling and PCR replicates are key to avoid obtaining false presence and absence and should be routinely corrected with the appropriate statistical tools referred to "site occupancy-detection modeling" (SODM) (Lahoz-Monfort et al., 2016). The SODM model shows precise estimation of the probability for the site occupancy, including overall probability of detection at sites where the species is present. The model provides unbiased estimation of occupancy when properly applied using large amount of initial data, even with a smaller amount of replications. Researchers (Ficetola et al., 2016) adopting SODM as part of their eDNA pipeline, give advice to avoid referring to single occurrences within one sample as reliable ones. Precautionary measures should be taken up before coming to conclusions that non-detection of species corresponds to species absence, and in converse that detections directly relies to species presence (Roussel et al., 2015) simply due to eDNA characteristics, such as potential longevity. In order to overcome the frontiers of eDNA techniques and to make it generally applicable within the SDM the above consistency is pivotal within the immense growing body of eDNA literature.

Combination of eDNA and SDMs

The method appears to be highly efficient on bony fish and amphibians with successful spatial representativeness in lotic and lentic systems (Civade et al., 2016). It has been shown that the eDNA samples are able to overcome spatial autocorrelation biases (Deiner et al., 2016) which are normally a result of conventional biodiversity assessments. eDNA seasonal diversity at the ecosystem scales (Bista et al., 2017) are key for more holistic understanding of the successful invasions of species within SDMs.

There are many possibilities of using eDNA for SDMs but currently one of the most important novel uses is a more precise sampling of absences which is sometimes difficult or impossible to obtain (Nezer et al., 2017). As commented, the information regarding species existence in certain system measured through eDNA can be susceptible to certain bias, due to eDNA characteristics. However, there exist approaches within the spatial modeling that might be applied to deal with the uncertainties from eDNA results. For instance, Dudík et al. (2006) presented the di-bias approach, which gives a higher weight in the models to those localities where presences or absences are more reliable. In the same way, those localities where eDNA is less reliable can receive a lower weight in the models, such weighting might correspond with the reported detection rates (**Table 1**). Therefore, there are possibilities from the SDMs to deal with the potential bias arising from using eDNA as a sampling technique which encourage its use despite current relative limitations. The ability to cope with the limitations and strength of the combination of these distinct research fields will benefit from the collaboration between molecular ecologists and modelers contributing to the evolution of two scientific disciplines (Coccia and Wang, 2016). Other disciplines apart from invasion ecology (e.g., biogeography or spatial ecology) might also benefit from future development of molecular ecology tools as a sampling technique. Thus, we highly recommend involving eDNA analysis into spatial models to predict future invasions and many other ecological processes. For example, targeting IAS hot spots and vectors of introduction, is a perfect starting point for detection of IAS and estimation of their future dispersal within the SDMs. Spatial representativeness of IAS within the SDMs is key to understanding the ecology behind their successful dispersal and the management of invasions.

CONCLUSION

Collaboration between modelers and molecular ecologists has a high potential to overcome the flaws of spatial distribution patterns due to difficulties or inconsistency in the information obtained through conventional surveys. The strength of the information that eDNA can provide is crucial as it fulfills the previously unidentified absences within the SDMs. The eDNA method is currently rapidly evolving and in the near future a mass of information related to IAS presence, absence as well as other species specific biological characteristic can be obtained and applied to, for example, mechanistic SDMs. Thus, its use is highly recommended with the aim of obtaining species distribution data for spatial models combining two scientific fields, useful as a helpful tool for IAS management and relevant policy requirements.

AUTHOR CONTRIBUTIONS

TM and MR-R compiled the knowledge based on their individual research and proposed the idea of eDNA methods usefulness within SDMs. They had both contributed to the written part of MS. MR was responsible for the table design and its content as well as a contribution to the overall MS. ET proposed the idea of eDNA and IAS, and revised the first draft of the paper.

ACKNOWLEDGMENTS

TM, MR-R, and MR are all Early Stage Researchers as part of Aquainvad- ED. We would like to thank for the full support to our supervisor Prof. Dr. Sonia Consuegra del Olmo. We would like to thank Prof. Frances E. Lucy for her generous contribution toward the English grammar correction. This work is a product of the Aquainvad-ED project and had received funding from the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement No 642197 (http://www.aquainvad-ed.com/).

REFERENCES

Balasingham, K. D., Walter, R. P., Mandrak, N. E., and Heath, D. D. (2017). Environmental DNA detection of rare and invasive fish species in two great lakes tributaries. *Mol. Ecol.* 26, 14395. doi: 10.1111/mec.14395

Barnes, M. A., and Turner, C. R. (2016). The ecology of environmental DNA and implications for conservation genetics. *Conserv. Genet.* 17, 1–17. doi: 10.1007/s10592-015-0775-4

Bista, I., Carvalho, G. R., Walsh, K., Seymour, M., Hajibabaei, M., Lallias, D., et al. (2017). Annual time-series analysis of aqueous eDNA reveals ecologically relevant dynamics of lake ecosystem biodiversity. *Nat. Commun.* 8:14087. doi: 10.1038/ncomms14087

Civade, R., Dejean, T., Valentini, A., Roset, N., Raymond, J.-C., Bonin, A., et al. (2016). Spatial representativeness of environmental DNA metabarcoding signal for fish biodiversity assessment in a natural freshwater system. *PLoS ONE* 11:e0157366. doi: 10.1371/journal.pone.0157366

Coccia, M., and Wang, L. (2016). Evolution and convergence of the patterns of international scientific collaboration. *Proc. Natl. Acad. Sci. U.S.A.* 113, 2057–2061. doi: 10.1073/pnas.1510820113

Deiner, K., Bik, H. M., Mächler, E., Seymour, M., Lacoursière-Roussel, A., Altermatt, F., et al. (2017a). Environmental DNA metabarcoding: transforming how we survey animal and plant communities. *Mol. Ecol.* 26, 5872–5895. doi: 10.1111/mec.14350

Deiner, K., Fronhofer, E. A., Mächler, E., Walser, J.-C., and Altermatt, F. (2016). Environmental DNA reveals that rivers are conveyer belts of biodiversity information. *Nat. Commun.* 7:12544. doi: 10.1038/ncomms12544

Deiner, K., Renshaw, M. A., Li, Y., Olds, B. P., Lodge, D. M., and Pfrender, M. E. (2017b). Long-range PCR allows sequencing of mitochondrial genomes from environmental DNA. *Methods Ecol. Evol.* doi: 10.1111/2041-210X.12836. [Epub ahead of print].

Dejean, T., Valentini, A., Duparc, A., Pellier-Cuit, S., Pompanon, F., Taberlet, P., et al. (2011). Persistence of environmental DNA in freshwater ecosystems. *PLoS ONE* 6:e23398. doi: 10.1371/journal.pone.0023398

Dejean, T., Valentini, A., Miquel, C., Taberlet, P., Bellemain, E., and Miaud, C. (2012). Improved detection of an alien invasive species through environmental DNA barcoding: the example of the American bullfrog *Lithobates catesbeianus*. *J. Appl. Ecol.* 49, 953–959. doi: 10.1111/j.1365-2664.2012.02171.x

Dougherty, M. M., Larson, E. R., Renshaw, M. A., Gantz, C. A., Egan, S. P., Erickson, D. M., et al. (2016). Environmental DNA (eDNA) detects the invasive rusty crayfish *Orconectes rusticus* at low abundances. *J. Appl. Ecol.* 53, 722–732. doi: 10.1111/1365-2664.12621

Dudík, M., Phillips, S. J., and Schapire, R. E. (2006). "Correcting sample selection bias in maximum entropy density estimation," in *Advances Neural Information Process Systems* (Vancouver), 323–330.

Dunker, K. J., Sepulveda, A. J., Massengill, R. L., Olsen, J. B., Russ, O. L., Wenburg, J. K., et al. (2016). Potential of environmental DNA to evaluate Northern pike (*Esox lucius*) eradication efforts: an experimental test and case study. *PLoS ONE* 11:e0162277. doi: 10.1371/journal.pone.0162277

Eichmiller, J. J., Best, S. E., and Sorensen, P.W. (2016). Effects of temperature and trophic state on degradation of environmental DNA in lake water. *Environ. Sci. Technol.* 50, 1859–1867. doi: 10.1021/acs.est.5b05672

Evans, N. T., Olds, B. P., Renshaw, M. A., Turner, C. R., Li, Y., Jerde, C. L., et al. (2016). Quantification of mesocosm fish and amphibian species diversity via environmental DNA metabarcoding. *Mol. Ecol. Resour.* 16, 29–41. doi: 10.1111/1755-0998.12433

Ficetola, G. F., Miaud, C., Pompanon, F., and Taberlet, P. (2008). Species detection using environmental DNA from water samples. *Biol. Lett.* 4, 423–425. doi: 10.1098/rsbl.2008.0118

Ficetola, G. F., Taberlet, P., and Coissac, E. (2016). How to limit false positives in environmental DNA and metabarcoding? *Mol. Ecol. Resour.* 16, 604–607. doi: 10.1111/1755-0998.12508

Fujiwara, A., Matsuhashi, S., Doi, H., Yamamoto, S., and Minamoto, T. (2016). Use of environmental DNA to survey the distribution of an invasive submerged plant in ponds. *Freshw. Sci.* 35, 748–754. doi: 10.1086/685882

Goldberg, C. S., Sepulveda, A., Ray, A., Baumgardt, J., and Waits, L. P. (2013). Environmental DNA as a new method for early detection of New Zealand mudsnails (*Potamopyrgus antipodarum*). *Freshw. Sci.* 32, 792–800. doi: 10.1899/13-046.1

Goldberg, C. S., Turner, C. R., Deiner, K., Klymus, K. E., Thomsen, P. F., Murphy, M. A., et al. (2016). Critical considerations for the application of environmental DNA methods to detect aquatic species. *Methods Ecol. Evol.* 7, 1299–1307. doi: 10.1111/2041-210X.12595

Goricki, Š., Stanković, D., Snoj, A., Kuntner, M., Jeffery, W. R., Trontelj, P., et al. (2017). Environmental DNA in subterranean biology: range extension and taxonomic implications for Proteus. *Sci. Rep.* 7:45054. doi: 10.1038/srep45054

Groom, Q. J., Adriaens, T., Desmet, P., Simpson, A., De Wever, A., Bazos, I., et al. (2017). Seven recommendations to make your invasive alien species data more useful. *Front. Appl. Mathemat. Stat.* 3:13. doi: 10.3389/fams.2017.00013

Hänfling, B., Lawson Handley, L., Read, D. S., Hahn, C., Li, J., Nichols, P., et al. (2016). Environmental DNA metabarcoding of lake fish communities reflects long-term data from established survey methods. *Mol. Ecol.* 25, 3101–3119. doi: 10.1111/mec.13660

Hinlo, R., Furlan, E., Suitor, L., and Gleeson, D. (2017). Environmental DNA monitoring and management of invasive fish: comparison of eDNA and fyke netting. *Management* 8, 89–100. doi: 10.3391/mbi.2017.8.1.09

Jerde, C. L., Chadderton, W. L., Mahon, A. R., Renshaw, M. A., Corush, J., Budny, M. L., et al. (2013). Detection of Asian carp DNA as part of a great lakes basin-wide surveillance program. *Can. J. Fish. Aquat. Sci.* 70, 522–526. doi: 10.1139/cjfas-2012-0478

Jerde, C. L., Mahon, A. R., Chadderton, W. L., and Lodge, D. M. (2011). "Sight-unseen" detection of rare aquatic species using environmental DNA. *Conserv. Lett.* 4, 150–157. doi: 10.1111/j.1755-263X.2010.00158.x

Jerde, C. L., Olds, B. P., Shogren, A. J., Andruszkiewicz, E. A., Mahon, A. R., Bolster, D., et al. (2016). Influence of stream bottom substrate on retention and transport of vertebrate environmental DNA. *Environ. Sci. Technol.* 50, 8770–8779. doi: 10.1021/acs.est.6b01761

Jiménez-Valverde, A., Peterson, A. T., Soberón, J., Overton, J. M., Aragón, P., and Lobo, J. M. (2011). Use of niche models in invasive species risk assessments. *Biol. Invasions* 13, 2785–2797. doi: 10.1007/s10530-011-9963-4

Lahoz-Monfort, J. J., Guillera-Arroita, G., and Tingley, R. (2016). Statistical approaches to account for false-positive errors in environmental DNA samples. *Mol. Ecol. Resour.* 16, 673–685. doi: 10.1111/1755-0998.12486

Larson, E. R., Renshaw, M. A., Gantz, C. A., Umek, J., Chandra, S., Lodge, D. M., et al. (2017). Environmental DNA (eDNA) detects the invasive crayfishes *Orconectes rusticus* and *Pacifastacus leniusculus* in large lakes of North America. *Hydrobiologia* 800, 173–185. doi: 10.1007/s10750-017-3210-7

Moyer, G. R., Díaz-Ferguson, E., Hill, J. E., and Shea, C. (2014). Assessing environmental DNA detection in controlled lentic systems. *PLoS ONE* 9:e103767. doi: 10.1371/journal.pone.0103767

Nathan, L. M., Simmons, M., Wegleitner, B. J., Jerde, C. L., and Mahon, A. R. (2014). Quantifying environmental DNA signals for aquatic invasive species across multiple detection platforms. *Environ. Sci. Technol.* 48, 12800–12806. doi: 10.1021/es5034052

Nezer, O., Bar-David, S., Gueta, T., and Carmel, Y. (2017). High-resolution species-distribution model based on systematic sampling and indirect observations. *Biodivers. Conserv.* 26, 421–437. doi: 10.1007/s10531-016-1251-2

Nichols, R. V., and Spong, G. (2017). An eDNA based SNP assay for ungulate species and sex identification. *Diversity* 9:33. doi: 10.3390/d9030033

Piaggio, A. J., Engeman, R. M., Hopken, M. W., Humphrey, J. S., Keacher, K. L., Bruce, W. E., et al. (2014). Detecting an elusive invasive species: a diagnostic PCR to detect Burmese python in Florida waters and an assessment of persistence of environmental DNA. *Mol. Ecol. Resour.* 14, 374–380. doi: 10.1111/1755-0998.12180

Roussel, J. M., Paillisson, J. M., Treguier, A., and Petit, E. (2015). The downside of eDNA as a survey tool in water bodies. *J. Appl. Ecol.* 52, 823–826. doi: 10.1111/1365-2664.12428

Schneider, J., Valentini, A., Dejean, T., Montarsi, F., Taberlet, P., Glaizot, O., et al. (2016). Detection of invasive mosquito vectors using environmental DNA (eDNA) from water samples. *PLoS ONE* 11:e0162493. doi: 10.1371/journal.pone.0162493

Secondi, J., Dejean, T., Valentini, A., Audebaud, B., and Miaud, C. (2016). Detection of a global aquatic invasive amphibian, *Xenopus laevis*, using environmental DNA. *Amphib. Reptil.* 37, 131–136. doi: 10.1163/15685381-00003036

Shokralla, S., Gibson, J., King, I., Baird, D., Janzen, D., Hallwachs, W., et al. (2016). Environmental DNA barcode sequence capture: targeted, PCR-free sequence capture for biodiversity analysis from bulk environmental samples. *BioRxiv* doi: 10.1101/087437

Sigsgaard, E. E., Nielsen, I. B., Bach, S. S., Lorenzen, E. D., Robinson, D. P., Knudsen, S. W., et al. (2016). Population characteristics of a large whale shark aggregation inferred from seawater environmental DNA. *Nat. Ecol. Evol.* 1:0004. doi: 10.1038/s41559-016-0004

Simmons, M., Tucker, A., Chadderton, W. L., Jerde, C. L., and Mahon, A. R. (2015). Active and passive environmental DNA surveillance of aquatic invasive species. *Can. J. Fish. Aquat. Sci.* 73, 76–83. doi: 10.1139/cjfas-2015-0262

Steel, E. A., Kennedy, M. C., Cunningham, P. G., and Stanovick, J. S. (2013). Applied statistics in ecology: common pitfalls and simple solutions. *Ecosphere* 4, 1–13. doi: 10.1890/ES13-00160.1

Strickler, K. M., Fremier, A. K., and Goldberg, C. S. (2015). Quantifying effects of UV-B, temperature, and pH on eDNA degradation in aquatic microcosms. *Biol. Conserv.* 183, 85–92. doi: 10.1016/j.biocon.2014.11.038

Taberlet, P., Coissac, E., Hajibabaei, M., and Rieseberg, L. H. (2012). Environmental DNA. *Mol. Ecol.* 21, 1789–1793. doi: 10.1111/j.1365-294X.2012.05542.x

Takahara, T., Minamoto, T., and Doi, H. (2013). Using environmental DNA to estimate the distribution of an invasive fish species in ponds. *PLoS ONE* 8:e56584. doi: 10.1371/journal.pone.0056584

Thomsen, P. F., and Willerslev, E. (2015). Environmental DNA–an emerging tool in conservation for monitoring past and present biodiversity. *Biol. Conserv.* 183, 4–18. doi: 10.1016/j.biocon.2014.11.019

Tréguier, A., Paillisson, J. M., Dejean, T., Valentini, A., Schlaepfer, M. A., and Roussel, J. M. (2014). Environmental DNA surveillance for invertebrate species: advantages and technical limitations to detect invasive crayfish *Procambarus clarkii* in freshwater ponds. *J. Appl. Ecol.* 51, 871–879. doi: 10.1111/1365-2664.12262

Tucker, A. J., Chadderton, W. L., Jerde, C. L., Renshaw, M. A., Uy, K., Gantz, C., et al. (2016). A sensitive environmental DNA (eDNA) assay leads to new insights on ruffe (*Gymnocephalus cernua*) spread in North America. *Biol. Invasions* 18, 3205–3222. doi: 10.1007/s10530-016-1209-z

Uchii, K., Doi, H., and Minamoto, T. (2016). A novel environmental DNA approach to quantify the cryptic invasion of non-native genotypes. *Mol. Ecol. Resour.* 16, 415–422. doi: 10.1111/1755-0998.12460

Václavík, T., and Meentemeyer, R. K. (2009). Invasive species distribution modeling (iSDM): are absence data and dispersal constraints needed to predict actual distributions? *Ecol. Modell.* 220, 3248–3258. doi: 10.1016/j.ecolmodel.2009.08.013

Wilcox, T. M., McKelvey, K. S., Young, M. K., Sepulveda, A. J., Shepard, B. B., Jane, S. F., et al. (2016). Understanding environmental DNA detection probabilities: a case study using a stream-dwelling char *Salvelinus fontinalis*. *Biol. Conserv.* 194, 209–216. doi: 10.1016/j.biocon.2015.12.023

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

The handling Editor declared a past co-authorship with one of the authors ET.

frontiers in Environmental Science

GENERAL COMMENTARY
published: 11 December 2017
doi: 10.3389/fenvs.2017.00085

Commentary: *Aedes albopictus* and *Aedes japonicus*—two invasive mosquito species with different temperature niches in Europe

Konstantinos Demertzis[1]*, *Lazaros Iliadis*[1] *and Vardis-Dimitrios Anezakis*[2]

[1] Department of Civil Engineering, Faculty of Mathematics Programming and General Courses, School of Engineering, Democritus University of Thrace, Xanthi, Greece, *[2] Department of Forestry and Management of the Environment and Natural Resources, Democritus University of Thrace, Orestiada, Greece*

Keywords: Asian bush mosquito, Asian tiger mosquito, climate change, invasive species, species distribution modeling, ensemble learning, machine learning

A commentary on

***Aedes albopictus* and *Aedes japonicus*—two invasive mosquito species with different temperature niches in Europe**
by Cunze, S., Koch, L. K., Kochmann, J., and Klimpel, S. Parasit. Vectors (2016). 9:573. doi: 10.1186/s13071-016-1853-2

OPEN ACCESS

Edited by:
Aristides (Aris) Moustakas,
Universiti Brunei Darussalam, Brunei

Reviewed by:
Dimitris Poursanidis,
Foundation for Research and Technology Hellas, Greece
Stelios Katsanevakis,
University of the Aegean, Greece

****Correspondence:***
Konstantinos Demertzis
kdemertz@fmenr.duth.gr

Specialty section:
This article was submitted to Environmental Informatics, a section of the journal Frontiers in Environmental Science

Received: *27 September 2017*
Accepted: *23 November 2017*
Published: *11 December 2017*

Citation:
Demertzis K, Iliadis L and Anezakis V-D (2017) Commentary: Aedes albopictus and Aedes japonicus—two invasive mosquito species with different temperature niches in Europe. Front. Environ. Sci. 5:85. doi: 10.3389/fenvs.2017.00085

INTRODUCTION

In this interesting and original study, the authors present an ensemble Machine Learning (ML) model for the prediction of the habitats' suitability, which is affected by the complex interactions between living conditions and survival-spreading climate factors. The research focuses in two of the most dangerous invasive mosquito species in Europe with the requirements' identification in temperature and rainfall conditions. Though it is an interesting approach, the ensemble ML model is not presented and discussed in sufficient detail and thus its performance and value as a tool for modeling the distribution of invasive species cannot be adequately evaluated.

METHODOLOGY USED

The authors use an Ensemble Approach (ENAP) based on 10 timely ML algorithms, aiming to draw up the habitats' maps for both species of mosquitoes. Ensemble methods are meta-algorithms that combine several techniques into a unique predictive model to decrease variance. For example, in Bagging different training data subsets are randomly drawn—with replacement—from the entire training dataset, to train a different classifier. In Boosting, resampling is strategically geared to provide the most informative training data for each consecutive classifier, or to improve predictions. Stacking, involves training to combine the predictions of several other learning algorithms (Zhou, 2012).

Unlike a statistical ensemble in statistical mechanics which is usually infinite, a ML ensemble consists of only a concrete finite set of alternative models, but typically allows for much more flexible structures to exist among those alternatives. Perhaps one of the earliest works on ensemble systems is the paper by Dasarathy and Sheela (1979). They first introduced an ENAP for partitioning the feature space, using two or more classifiers, in a divide-and-conquer fashion. Over a decade later, Hansen and Salamon (1990) showed the variance reduction property of an ENAP.

They managed to improve the generalization performance of an ANN by using an ensemble of similarly configured ANN. But it was Schapire's work that has put the ENAP at the center of ML research, as he has proven that a strong classifier can be generated by combining weak classifiers (Schapire, 1990). Finally, Buisson et al. suggested that attention should be paid to the use of predictions ensembles resulting from the application of several statistical methods. Forecasted impacts should always be provided with an assessment of their uncertainty (Buisson et al., 2010).

Unfortunately, the authors of this interesting paper, do not offer a deep description of the proposed ENAP and it is not clear if their approach can cover the main points of the ensemble techniques. For example, the proposed ENAP convert species' probability of occurrence into binary presence-absence data using a predefined threshold. Assessing models based on presence only data, it is difficult to learn the overall species occurrence probability, based on false or misleading information or unjustified simplifying assumptions, because there is typically no validation data with true presences and absences (Hastie and Fithian, 2013). The ENAP that was proposed cannot surmount this problem, it only makes it more hidden.

ALIEN SPECIES DISTRIBUTION MODELING AND MACHINE LEARNING ENSEMBLES MODELS

Current practices in Alien Species Distribution Modeling (ASDM) algorithms (Lorena et al., 2011; Duan et al., 2014; Shabani et al., 2016), include Profile Methods (BIOCLIM, ENFA) (Lorena et al., 2011; Duan et al., 2014; Shabani et al., 2016), Regression-based techniques (GLM, MARS) (Lorena et al., 2011; Duan et al., 2014; Shabani et al., 2016), ML techniques (MAXENT, ANN, SVM) (Lorena et al., 2011; Duan et al., 2014; Shabani et al., 2016).

A widely used and effective method in ASDM involves creating ML ensembles' models (Duan et al., 2014). The two most important advantages of ENAP focus on the fact that they offer better prediction and more stable and robust models, as the overall behavior of a multiple model is less noisy than a corresponding single one (Kuncheva, 2004; Zhou, 2012). For example, in Zhang and Zhang (2012) the authors propose an effective ENAP to assess the impacts of predictor variables and ASDM. In Daliakopoulos et al. (2017) the Random Forest EANP has proven that it can provide a better understanding of facilitating and limiting factors of alien species presence, both for research and management purposes. Finally, Lauzeral et al. (2012) proposes an iterative ENAP to ensure noise absence and hence to improve the predictive reliability of ensemble modeling of species distributions.

Some of the most important points related to the operation and use of the ENAP that should be included and discussed thoroughly by the authors are presented below:

1. The ensemble size of the proposed model. The number of classifiers included in the creation of an ensemble model has a large impact on the accuracy of the prediction (Kuncheva, 2004; Zhou, 2012). Regarding the proposed ENAP, a 10 state-of-the-art algorithms used, nevertheless without thorough analysis and explanation. On the other hand, their theoretical framework of Ensemble Learning shows that using the same number of independent component classifiers as class labels gives the highest accuracy (Hamed and Can, 2016).
2. A detailed and complete description and justification of the classifiers selection. The choice of the proper classifiers (e.g., ANN) to be included in an ENAP (Kuncheva, 2004; Zhou, 2012) should be based on the selection of the implementation mode and on the parameters' settings which can lead to different decision boundaries, even if all other parameters remain constant (Kuncheva, 2004; Zhou, 2012). It is a fact that there is no point or advantage to combining a group of models that are identical and generalize in the same way (López et al., 2007; Bougoudis et al., 2014). In the proposed ENAP, both GLM and MAXENT were used, and there is no clear explanation on how the authors have chosen this specific architecture. As shown by Renner and Warton (2013) MAXENT is equivalent to a GLM with a Poisson error structure and differing only in the intercept term, which is scale-dependent in MAXENT. One cannot argue that MAXENT has different predictive performance than a GLM when they are equivalent.
3. A clear and sufficiently detailed discussion-explanation on the determination and handling of the weights employed by the distinct ensemble models (Kuncheva, 2004; Zhou, 2012). The weight vector is a very important parameter in the process of training an ENAP, as it is used in the determination of the classifiers' performance and of the classification confidence level (Kuncheva, 2004; Zhou, 2012). The authors do not include a detailed description of the weights employed by the distinct ensemble models, with no attempt to tie them to the problem at hand.
4. Clear description of the process that has determined the optimal model, its potential hybrid nature and justification of the proposed ensemble's architecture reliability. This can be done using inclusion of diagrams or algorithms. The variance of prediction results in a ML model is one of the most important measures for assessing the credibility of the method (Kuncheva, 2004; Zhou, 2012). The work by Yackulic et al. (2013) shows that MAXENT model outputs (i.e., maps) are presented completely casually and without providing readers with any means to critically examine modeled relationships. This fact may be hidden or masked within proposed ENAP, but the problem remains.

DISCUSSION AND CONCLUSIONS

It is worth noting that in general an ENAP can lead to much better prediction results, while offering generalization. This is one of the key issues in the field of ML, as it can reduce bias and variance and it has the potential to eliminate overfitting (Kuncheva, 2004; Zhou, 2012). Moreover, it implements robust predictive models capable of responding to high complexity problems such as those of spreading invasive species (Demertzis and Iliadis, 2015, 2017). However, the development of these models should not be done in a black box mode research and it should be accompanied by a set of in-depth analysis regarding key training and operation decision points, thus allowing critical readers to fully and thoroughly evaluate the proposed methodology and to promote research in the broader scope. Finally, there are cases where wide variety of comparatively model-free forecasting methods outperforms the correct mechanistic data-driven model. However, according to Moustakas (2017) "*if one simply relies on data-driven science, several components of scientific methods will be made poorer.*"

AUTHOR CONTRIBUTIONS

KD and LI conceived of the presented idea. KD and LI developed the theoretical formalism, performed the analytic calculations and performed the numerical simulations. KD, LI, and V-DA verified the analytical methods. All the authors contributed to the final version of the manuscript. LI supervised the project.

REFERENCES

Bougoudis, I., Iliadis, L., and Papaleonidas, A. (2014). "Fuzzy inference ANN ensembles for air pollutants modeling in a major urban area: the case of athens," in *Proceedings of the 15th Engineering Applications of Neural Networks*, *Vol. 459* (Cham: Springer; Communications in Computer and Information Science), 1–14.

Buisson, L., Thuiller, W., Casajus, N., Lek, S., and Grenouillet, G. (2010). Uncertainty in ensemble forecasting of species distribution. *Global Change Biol.* 16, 1145–1157. doi: 10.1111/j.1365-2486.2009.02000.x

Daliakopoulos, I. N., Katsanevakis, S., and Moustakas, A. (2017). Spatial downscaling of alien species presences using machine learning. *Front. Earth Sci.* 5:60. doi: 10.3389/feart.2017.00060

Dasarathy, B. V., and Sheela, B. V. (1979). Composite classifier system design: concepts and me-thodology. *Proc. IEEE* 67, 708–713. doi: 10.1109/PROC.1979.11321

Demertzis, K., and Iliadis, L. (2015). "Intelligent bio-inspired detection of food borne pathogen by DNA barcodes: the case of invasive fish species lagocephalus sceleratus," in *Engineering Applications of Neural Networks. Communications in Computer and Information Science, Vol. 517,* eds L. Iliadis and C. Jayne (Cham: Springer), 89–99.

Demertzis, K., and Iliadis, L. (2017). "Adaptive elitist differential evolution extreme learning machines on big data: intelligent recognition of invasive species," in *Advances in Big Data; Advances in Intelligent Systems and Computing, Vol. 529*, eds P. Angelov, Y. Manolopoulos, L. Iliadis, A. Roy, and M. Vellasco (Cham: Springer), 333–345.

Duan, R.-Y., Kong, X.-Q., Huang, M.-Y., Fan, W.-Y., and Wang, Z.-G. (2014). The predictive performance and stability of six species distribution models. *PLoS ONE* 9:e112764. doi: 10.1371/journal.pone.0112764

Hamed, R. B., and Can, F. (2016). "A theoretical framework on the ideal number of classifiers for online ensembles in data streams," in *CIKM* (Indianapolis, IN: ACM), 2053.

Hansen, L. K., and Salamon, P. (1990). Neural network ensembles. *IEEE Trans. Pattern Anal. Mach. Intellig.* 12, 993–1001. doi: 10.1109/34.58871

Hastie, T., and Fithian, W. (2013). Inference from presence-only data; the ongoing controversy. *Ecography* 36, 864–867. doi: 10.1111/j.1600-0587.2013.00321.x

Kuncheva, L. (2004). *Combining Pattern Classifiers: Methods and Algorithms.* Hoboken, NJ: Wiley.

Lauzeral, C., Grenouillet, G., and Brosse, S. (2012). Dealing with noisy absences to optimize species distribution models: an iterative ensemble modelling approach. *PLoS ONE* 7:e49508. doi: 10.1371/journal.pone.0049508

López, M., Melin, P., and Castillo, O. (2007). A method for creating ensemble neural networks using a sampling data approach. *Ther. Adv. Appl. Fuzzy Logic* 42, 772–780. doi: 10.1007/978-3-540-72434-6_78

Lorena, A. C., Jacintho, L. F. O., Siqueira, M. F., De Giovanni, R., Lohmann, L. G., de Carvalho, A. C. P. L. F., et al. (2011). Comparing machine learning classifiers in potential distribution modelling. *Expert Syst. Appl.* 38, 5268–5275. doi: 10.1016/j.eswa.2010.10.031

Moustakas, A. (2017). Spatio-temporal data mining in ecological and veterinary epidemiology. *Stochast. Environ. Res. Risk Assess.* 31, 829–834. doi: 10.1007/s00477-016-1374-8

Renner, I. W., and Warton, D. I. (2013). Equivalence of MAXENT and poisson point process models for species distribution modeling in ecology. *Biometrics* 69, 274–281. doi: 10.1111/j.1541-0420.2012.01824.x

Schapire, R. E. (1990). The strength of weak learnability. *Mach. Learn.* 5, 197–227.

Shabani, F., Kumar, L., and Ahmadi, M. (2016). A comparison of absolute performance of different correlative and mechanistic species distribution models in an independent area. *Ecol. Evol.* 6, 5973–5986. doi: 10.1002/ece3.2332

Yackulic, C. B., Chandler, R., Zipkin, E. F., Royle, J. A., Nichols, J. D., Grant, E. H. C., et al. (2013). Presence-only modelling using MAXENT: when can we trust the inferences? *Methods Ecol. Evol.* 4, 236–243. doi: 10.1111/2041-210x.12004

Zhang, Q., and Zhang, X. (2012). Impacts of predictor variables and species models on simulating *Tamarix ramosissima* distribution in Tarim Basin, northwestern China. *J. Plant Ecol.* 5, 337–345. doi: 10.1093/jpe/rtr049

Zhou, Z.-H. (2012). *Ensemble Methods: Foundations and Algorithms.* London: CRC Press.

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers in Marine Science

OPINION
published: 09 February 2018
doi: 10.3389/fmars.2018.00038

Uncertainty in Marine Invasion Science

Stelios Katsanevakis[1] and Aristides Moustakas[2]*

[1] Department of Marine Sciences, University of the Aegean, Mytilene, Greece, [2] Institute for Applied Data Analytics, Universiti Brunei Darussalam, Bandar Seri Begawan, Brunei

Keywords: alien species, invasive species, marine environment, uncertainty analysis, biological invasions

OPEN ACCESS

Edited by:
Juergen Pilz,
Alpen-Adria-Universität Klagenfurt,
Austria

Reviewed by:
Laura Uusitalo,
Finnish Environment Institute (SYKE),
Finland

****Correspondence:***
Stelios Katsanevakis
stelios@katsanevakis.com

Specialty section:
This article was submitted to
Environmental Informatics,
a section of the journal
Frontiers in Marine Science

Received: *28 September 2017*
Accepted: *25 January 2018*
Published: *09 February 2018*

Citation:
Katsanevakis S and Moustakas A
(2018) Uncertainty in Marine Invasion
Science. Front. Mar. Sci. 5:38.
doi: 10.3389/fmars.2018.00038

Uncertainty can be considered as an attribute of (or reciprocally lack of) information (Zadeh, 2005). Nevertheless, this vital attribute is usually missing from marine invasion science studies (**Figure 1**), despite the fact that methods such as Monte Carlo simulations, sensitivity analysis, Bayesian uncertainty assessment, and Latin hypercube sampling appear more frequently in the ecological literature (e.g., Harwood and Stokes, 2003; Moustakas and Evans, 2015; Aiello-Lammens and Akçakaya, 2017). Most studies of marine invasions are plagued by uncertainties, which in most cases are totally ignored or if acknowledged are not properly addressed or quantified. Herein we provide some examples of uncertainties in invasion science, aiming to highlight the existing gaps and stress the need for the development and implementation of frameworks, methods and tools that will assist invasion biologists to adequately tackle uncertainty.

UNCERTAINTY IN ALIEN STATUS

Although there are various definitions of alien species (Falk-Petersen et al., 2006), the most widely accepted one is that of species, which, by human agency, have managed to overcome physical barriers and colonize new regions beyond their natural range. In a rapidly changing world, it is often difficult to attribute a new arrival to a region to human activities. Species' natural ranges are inherently dynamic, shaped by natural variability, and thus sometimes it is impossible to determine the relative importance of natural and anthropogenic factors in distributional changes. Furthermore, assessing the alien or native status of rare species or of very old invasions is often impossible due to the lack of historical data. For example, shipping has been acting as an invasion pathway for many millennia making it difficult to evaluate the biogeographic status of species introduced to new marine regions hundreds or thousands years ago, some of which may have gradually become cosmopolitan. The term "cryptogenic" has been used by invasion biologists (Carlton, 1996) to indicate species that could be either native or alien but for which uncertainty is high. Assessing the native/alien status of species is important both for ecological science and management, hence a standardized framework with robust definitions, transparent criteria for classification and proper uncertainty assessment is urgently needed.

UNCERTAINTY IN ALIEN SPECIES INVENTORIES

National or regional inventories of alien species are a valuable tool for both invasion science and management, especially for prioritizing pathways, developing prevention or control strategies, and assessing (through trend analyses) of the effectiveness of management measures. However, alien species inventories suffer from a number of uncertainties e.g. in species identification (taxonomic uncertainty), incomprehensive search of data sources, low spatial resolution, poor documentation of data and knowledge, and inadequate native range information (McGeoch et al., 2012). This has resulted in regular revisions of such inventories, excluding species previously included (e.g., Zenetos et al., 2017), and even in scientific disputes among research groups

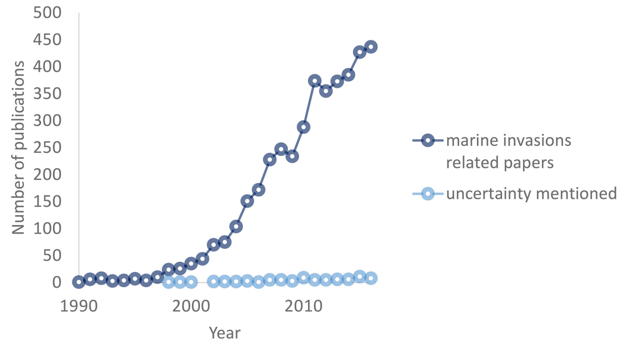


FIGURE 1 | Time series of publications registered in Scopus, including in the title or abstract or keywords at least one the words "biological invasions" or "non-indigenous species" or "alien species" or "invasive species," and at least one of the words "marine" or "sea" (dark blue markers). If in addition one of the words "uncertainty" or "confidence" is included (light blue markers) the number of publications drops to 2% of the initial number (91/4415).

(e.g., Galil, 2009, 2012; Zenetos, 2010; Zenetos et al., 2017). To avoid misuse of inventories, misunderstandings, mistrust and unnecessary conflicts, proper account of the level of confidence and inherent uncertainties should be included in all species inventories. The uncertainty in inventories is likely to increase with the use of citizen science (ElQadi et al., 2017) as well as with mining (geo-referenced) social media (Daume, 2016). While these techniques are likely to increase the total amount of data availability, the inclusion of non-experts in the process of data collection as well as the automated way of classification is likely to increase uncertainty as well as to introduce noise and spurious correlations in the dataset (Moustakas, 2017). In addition, as more datasets become publicly accessible it is likely that datasets collected under different protocols or for different purposes may be merged together and thus introduce additive uncertainties (Evans and Moustakas, 2016).

UNCERTAINTY IN PATHWAY ASSESSMENT

Assessment of the pathways of introduction of alien species (i.e., any means that allows the entry or spread of an alien species into a new region) is fundamental to biological invasion risk assessments, management of invasive species, monitoring and surveillance of new introductions (Essl et al., 2015). However, assigning each marine alien introduction to a specific pathway is subject to substantial uncertainty (Zenetos et al., 2012). Apart from intentional introductions, in most other cases more than one possible pathway can be inferred based on the human activities in or near the locality of first record. However, such inferences are largely based on expert judgement and suffer from lack of evidence. For example, the invasive crab *Percnon gibbesi* was recorded for the first time in the Mediterranean Sea in 1999 in many different locations, and its introduction has been attributed to a number of different possible pathways by different authors: shipping, aquarium trade or larval drift by the Atlantic surface current entering the Mediterranean (see (Katsanevakis et al., 2011) and references therein); in the latter case the species should not be considered as an alien, because its introduction was not mediated by humans but happened through natural processes. Many other possible pathways that are commonly neglected or understudied, e.g., aquarium trade (Padilla and Williams, 2004; Strecker et al., 2011) or marine litter (Barnes, 2002), could have had much higher importance than currently acknowledged. It is vital that such uncertainties in pathway assessments are transparent by providing an estimate of the related confidence and highlight possibilities for alternative pathways (Zenetos et al., 2012; Essl et al., 2015).

UNCERTAINTY IN IMPACT AND RISK ASSESSMENTS

Impact assessments depend on uncertain data such as the actual distribution of alien species, the vulnerability of ecosystems and their services, heterogeneity of biophysical processes, and spatiotemporal variability in the magnitude of impacts. Assessments of cumulative impacts of many invasive species to marine ecosystems suffer from additional uncertainties related to insufficient data, sensitivity weights, type of responses of ecosystems to invasive species, type of multiple species effects (additive or with synergistic or antagonistic interactions), and resolution of spatial data (Katsanevakis et al., 2016). Such uncertainties can substantially affect impact assessments and jeopardize their quality and consequently their value for management. Important gaps exist in properly addressing uncertainty in impact assessments, which, if at all, are restricted in qualitative accounts based largely on expert judgement (Blackburn et al., 2014). The International Pest Risk Mapping Workgroup has identified "improving representations of uncertainty" among the most important actions to improve pest risk assessment procedures (Venette et al., 2010). Research effort is needed to further develop proper tools that will allow addressing uncertainty in impact and risk assessments in an adequate and transparent way.

UNCERTAINTY IN FIELD SURVEYS

Monitoring of marine aliens species is based on field surveys, commonly conducted by visual surveys through SCUBA diving or the use of sampling and fishing gears. One important source of error in field surveys of marine species is imperfect detectability or imperfect catchability, i.e., the fact that not all individuals or all target species in a study area can be detected by observers (divers) or caught by sampling devices or fishing gears (Katsanevakis et al., 2012). Although many methods have been developed that properly account for imperfect detectability/catchability, such as distance sampling, mark-recapture, repeated presence-absence surveys for occupancy estimation (Issaris et al., 2012; Katsanevakis et al., 2012), the number of marine field surveys that apply such methods remains very limited. Failure to properly account for detectability leads to underestimation of the population state variable (e.g., abundance, population density, or occupancy), or even total failure to detect an alien species and thus underestimation of alien species richness in community studies. This is further intensified by the inability to properly identify alien species *in situ* in visual surveys. To improve the performance of field surveys and reduce uncertainties it is imperative to select an adequate method properly accounting for detectability, and apply field protocols and tools (sampling devices, photos, video) that will reduce identification uncertainties.

UNCERTAINTY IN DISTRIBUTION MODELING

Species distribution models (SDM) have been extensively used to predict the potential present or future distribution of marine alien species, based on a limited set of observations and a set of environmental variables that presumably describe their niches (Peterson, 2003). Uncertainty arising due to the applied modeling technique, spatial resolutions, scales, data availability, climate change and subsequent biological responses, model selection and evaluation methods can be substantially high (Beale and Lennon, 2012). SDMs are based in the (silent) presupposition that there exists information in the species distribution that can be used for estimating the species' niche, though there is no certainty that all niche axes have limits within the spatial extent of the dataset (Beale and Lennon, 2012). In addition, a critical form of uncertainty in SDMs is identifying the dimensions of explanatory variables (Beale et al., 2010); omitting informative variables produces poor model predictions while including uninformative variables may result in correlations with informative variables and thus result in reduced parameter estimation accuracy (Beale et al., 2010). Improved assessment of errors and uncertainties is among the prevailing challenges facing SDM research (Guisan and Thuiller, 2005).

CONCLUDING REMARKS

Quantifying or at least acknowledging uncertainty is an elementary exercise in science, as without uncertainty assessment it is hard to envisage future improvements (see e.g., Benali et al., 2017). In addition, biological invasions and their impacts is a very timely topic receiving high public attention. If scientists are perceived by the public either to overstate their findings in order to receive high visibility or to downplay the uncertainty of their findings, society is likely to lose confidence in the outputs of invasion science.

It is evident from the above non-exhaustive list that invasion science is challenged by various sources of uncertainty. Such uncertainties, if ignored, render many types of analyses and results doubtful and of limited practical use to policy makers and marine managers. To further advance invasion science in the marine environment we need (1) to acknowledge the need for appropriate assessments of uncertainty in all aspects of biological invasions research, (2) put research effort in improving/developing methods and tools for uncertainty analyses, (3) report uncertainties and their effects on research outcomes in all scientific outputs and technical reports.

AUTHOR CONTRIBUTIONS

All authors listed have made a substantial, direct and intellectual contribution to the work, and approved it for publication.

REFERENCES

Aiello-Lammens, M. E., and Akçakaya, H. R. (2017). Using global sensitivity analysis of demographic models for ecological impact assessment. *Conserv. Biol.* 31, 116–125. doi: 10.1111/cobi.12726

Barnes, D. K. A. (2002). Invasions by marine life on plastic debris. *Nature* 416, 808–809. doi: 10.1038/416808a

Beale, C. M., and Lennon, J. J. (2012). Incorporating uncertainty in predictive species distribution modelling. *Philos. T. Roy. Soc. B* 367, 247–258. doi: 10.1098/rstb.2011.0178

Beale, C. M., Lennon, J. J., Yearsley, J. M., Brewer, M. J., and Elston, D. A. (2010). Regression analysis of spatial data. *Ecol. Lett.* 13, 246–264. doi: 10.1111/j.1461-0248.2009.01422.x

Benali, A., Sa, A. C. L., Ervilha, A. R., Trigo, R. M., Fernandes, P. M., and Pereira, J. M. C. (2017). Fire spread predictions: sweeping uncertainty under the rug. *Sci. Total Environ.* 592, 187–196. doi: 10.1016/j.scitotenv.2017.03.106

Blackburn, T. M., Essl, F., Evans, T., Hulme, P. E., Jeschke, J. M., Kühn, I., et al. (2014). A unified classification of alien species based on the magnitude of their environmental impacts. *PLoS Biol.* 12:e1001850. doi: 10.1371/journal.pbio.1001850

Carlton, J.T. (1996). Biological invasions and cryptogenic species. *Ecology* 77, 1653–1655. doi: 10.2307/2265767

Daume, S. (2016). Mining Twitter to monitor invasive alien species — An analytical framework and sample information topologies. *Ecol. Inform.* 31, 70–82. doi: 10.1016/j.ecoinf.2015.11.014

ElQadi, M. M., Dorin, A., Dyer, A., Burd, M., Bukovac, Z., and Shrestha, M. (2017). Mapping species distributions with social media geo-tagged images: case studies of bees and flowering plants in Australia. *Ecol. Inform.* 39, 23–31. doi: 10.1016/j.ecoinf.2017.02.006

Essl, F., Bacher, S., Blackburn, T., Booy, O., Brundu, G., Brunel, S. et al. (2015). Crossing frontiers in tackling pathways of biological invasions. *BioScience* 65, 769–782. doi: 10.1093/biosci/biv082

Evans, M. R., and Moustakas, A. (2016). A comparison between data requirements and availability for calibrating predictive ecological models for lowland UK woodlands: learning new tricks from old trees. *Ecol. Evol.* 6, 4812–4822. doi: 10.1002/ece3.2217

Falk-Petersen, J., Bohn, T., and Sandlund, O. T. (2006). On the numerous concepts in invasion biology. *Biol. Invasions* 8, 1409–1424. doi: 10.1007/s10530-005-0710-6

Galil, B. S. (2009). Taking stock: inventory of alien species in the Mediterranean Sea. *Biol. Invasions* 11, 359–372. doi: 10.1007/s10530-008-9253-y

Galil, B. S. (2012). Truth and consequences: the bioinvasion of the Mediterranean Sea. *Integr. Zool.* 7, 299–311. doi: 10.1111/j.1749-4877.2012.00307.x

Guisan, A., and Thuiller, W. (2005). Predicting species distribution: offering more than simple habitat models. *Ecol. Lett.* 8, 993–1009. doi: 10.1111/j.1461-0248.2005.00792.x

Harwood, J., and Stokes, K. (2003). Coping with uncertainty in ecological advice: lessons from fisheries. *Trends Ecol. Evol.* 18, 617–622. doi: 10.1016/j.tree.2003.08.001

Issaris, Y., Katsanevakis, S., Salomidi, M., Tsiamis, K., Katsiaras, N., and Verriopoulos, G. (2012). Occupancy estimation of marine demersal species: dealing with imperfect detectability. *Mar. Ecol. Prog. Ser.* 453, 95–106. doi: 10.3354/meps09668

Katsanevakis, S., Poursanidis, D., Yokes, B., Mačić, V., Beqiraj, S., Kashta, L., et al. (2011). Twelve years after the first report of the crab *Percnon gibbesi* (H. Milne Edwards, 1853) in the Mediterranean: current distribution and invasion rates. *J. Biol. Res*. 16, 224–236.

Katsanevakis, S., Tempera, F., and Teixeira, H. (2016). Mapping the impact of alien species on marine ecosystems: the Mediterranean Sea case study. *Divers. Distrib*. 22, 694–707. doi: 10.1111/ddi.12429

Katsanevakis, S., Weber, A., Pipitone, C., Leopold, M., Cronin, M., Scheidat, M., et al. (2012). Monitoring marine populations and communities: review of methods and tools dealing with imperfect detectability. *Aquat. Biol*. 16, 31–52. doi: 10.3354/ab00426

McGeoch, M. A., Spear, D., Kleynhans, E. J., and Marais, E. (2012). Uncertainty in invasive species listing. *Ecol. Appl.* 22, 959–971. doi: 10.1890/11-1252.1

Moustakas, A. (2017). Spatio-temporal data mining in ecological and veterinary epidemiology. *Stoch. Environ. Res. Risk Assess.* 31, 829–834. doi: 10.1007/s00477-016-1374-8

Moustakas, A., and Evans, M. (2015). Coupling models of cattle and farms with models of badgers for predicting the dynamics of bovine tuberculosis (TB). *Stoch. Environ. Res. Risk Assess.* 29, 623–635. doi: 10.1007/s00477-014-1016-y

Padilla, D. K., and Williams, S. L. (2004). Beyond ballast water: aquarium and ornamental trades as sources of invasive species in aquatic ecosystems. *Front. Ecol. Environ.* 2, 131–138. doi: 10.1890/1540-9295(2004)002[0131:BBWAAO]2.0.CO;2

Peterson, A.T. (2003). Predicting the geography of species' invasions via ecological niche modeling. *Q. Rev. Biol.* 78, 419–433. doi: 10.1086/378926

Strecker, A. L., Campbell, P. M., and Olden, J. D. (2011). The aquarium trade as an invasion pathway in the Pacific Northwest. *Fisheries* 36, 74–85. doi: 10.1577/03632415.2011.10389070

Venette, R. C., Kriticos, D. J., Magarey, R., Koch, F., Baker, R. H. A., Worner, S. P., et al. (2010). Pest risk maps for invasive alien species: a roadmap for improvement. *Bioscience* 80, 349–362. doi: 10.1525/bio.2010.60.5.5

Zadeh, L. A. (2005). Toward a generalized theory of uncertainty (GTU)—an outline. *Inform. Sci.* 172, 1–40. doi: 10.1016/j.ins.2005.01.017

Zenetos, A. (2010). Trend in aliens species in the Mediterranean. An answer to Galil, 2009 ≪Taking stock: inventory of alien species in the Mediterranean Sea≫. *Biol. Invasions* 12, 3379–3381. doi: 10.1007/s10530-009-9679-x

Zenetos, A., Gofas, S., Morri, C., Rosso, A., Violanti, D., et al. (2012). Alien species in the Mediterranean Sea by 2012. A contribution to the application of European Union's Marine Strategy Framework Directive (MSFD). Part 2. Introduction trends and pathways. *Medit. Mar. Sci.* 13, 328–352. doi: 10.12681/mms.327

Zenetos, A., Çinar, M. E., Crocetta, F., Golani, D., Rosso, A., Servello, G. et al. (2017). Uncertainties and validation of alien species catalogues: the Mediterranean as an example. *Est. Coast. Shelf Sci.* 191, 171–187. doi: 10.1016/j.ecss.2017.03.031

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Advantages of publishing in Frontiers

OPEN ACCESS
Articles are free to read, for greatest visibility

COLLABORATIVE PEER-REVIEW
Designed to be rigorous – yet also collaborative, fair and constructive

FAST PUBLICATION
Average 85 days from submission to publication (across all journals)

COPYRIGHT TO AUTHORS
No limit to article distribution and re-use

TRANSPARENT
Editors and reviewers acknowledged by name on published articles

SUPPORT
By our Swiss-based editorial team

IMPACT METRICS
Advanced metrics track your article's impact

GLOBAL SPREAD
5'100'000+ monthly article views and downloads

LOOP RESEARCH NETWORK
Our network increases readership for your article

Frontiers
EPFL Innovation Park, Building I • 1015 Lausanne • Switzerland
Tel +41 21 510 17 00 • Fax +41 21 510 17 01 • info@frontiersin.org
www.frontiersin.org

Find us on