



OPEN ACCESS

EDITED BY

Feng Chen,
Dallas County, United States

REVIEWED BY

Patrick Wagnon,
UMR5001 Institut des Géosciences de
l'Environnement (IGE), France
Xujiang Zhao,
NEC Laboratories America Inc, United States

*CORRESPONDENCE

Hamish Steptoe

✉ hamish.steptoe@metoffice.gov.uk

RECEIVED 31 March 2023

ACCEPTED 21 July 2023

PUBLISHED 08 August 2023

CITATION

Steptoe H and Economou T (2023) Proliferation of atmospheric datasets can hinder policy making: a data blending technique offers a solution. *Front. Big Data* 6:1198097. doi: 10.3389/fdata.2023.1198097

COPYRIGHT

Crown Copyright © 2023 Met Office. Authors: Steptoe and Economou. This is an open-access article distributed under the terms of the [Creative Commons Attribution License \(CC BY\)](https://creativecommons.org/licenses/by/4.0/). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

Proliferation of atmospheric datasets can hinder policy making: a data blending technique offers a solution

Hamish Steptoe^{1*} and Theo Economou²

¹Met Office, Exeter, United Kingdom, ²Climate and Atmosphere Research Centre (CARE-C), The Cyprus Institute, Nicosia, Cyprus

The proliferation of atmospheric datasets is a key outcome from the continued development and advancement of our collective scientific understanding. Yet often datasets describing ostensibly identical processes or atmospheric variables provide widely varying results. As an example, we analyze several datasets representing rainfall over Nepal. We show that estimates of extreme rainfall are highly variable depending on which dataset you choose to look at. This leads to confusion and inaction from policy-focused decision makers. Scientifically, we should use datasets that sample a range of creation methodologies and prioritize the use of data science techniques that have the flexibility to incorporate these multiple sources of data. We demonstrate the use of a statistically interpretable data blending technique to help discern and communicate a consensus result, rather than imposing a priori judgment on the choice of dataset, for the benefit of policy decision making.

KEYWORDS

data science, decision making, data blending, precipitation, Nepal, Generalized Additive Models, statistics

1. Introduction

Data underpins all scientific analysis. But bridging the gap from science to policy making typically requires a consistent scientific message; uncertainty or a lack of consensus is frequently used to justify inaction (Mccright and Dunlap, 2000; Orlove et al., 2020). In atmospheric science, the availability of data on which to conduct scientific analysis is considerable and reflects the wide variety of data collection and processing techniques. To demonstrate the variation of results that could be obtained from common, open access datasets, we examine summer rainfall extremes in Nepal.

In Nepal, the South Asian summer monsoon season (June to September, JJAS), contributes 70–80% of annual rainfall totals (DHM, 2022). Extreme precipitation events that occur in the monsoon season have wide-ranging impacts including flooding and landslides, which can be damaging and costly to a variety of infrastructure. Within the context of hydropower, damage to infrastructure as a result of climate induced hazards, is most often associated with extreme rainfall accumulation occurring in the monsoon season (Basnyat and Watkiss, 2017). Therefore, planning new infrastructure projects, such as hydropower plants or urban development, should incorporate an understanding of the likelihood of extreme rainfall events.

2. Methods

The key requirements for our data blending framework is the ability to: (i) be spatially and temporally consistent, (ii) account for data hierarchy, (iii) integrate unobserved uncertainty that accounts for the variability between datasets and (iv) flexibly characterize the variability in the data via a range of possible probability distributions. We construct our data blending framework based on Generalized Additive Models (GAMs) (Hastie and Tibshirani, 1990; Wood, 2017). For this case study, we demonstrate extreme value analysis of RX1day JJAS block-maxima of extreme precipitation, modeled using the Generalized Extreme Value (GEV) distribution:

$$\begin{aligned}
 Y_{s,t,m} &\sim \text{GEV}(\mu_{s,t,m}, \sigma_{s,t,m}, \xi_m) \\
 \mu_{s,t,m} &= \beta_0 + \mathcal{F}(\text{year}_t) + \mathcal{G}(\text{lon}_s, \text{lat}_s) \\
 &\quad + \mathcal{H}(\text{lon}_s, \text{lat}_s, u_m^{(\mu)}) + \epsilon_i \\
 \log(\sigma_{s,t,m}) &= \gamma_0 + \mathcal{F}(\text{year}(t)) + \mathcal{G}(\text{lon}_s, \text{lat}_s) \\
 &\quad + \mathcal{H}(\text{lon}_s, \text{lat}_s, u_m^{(\sigma)}) + \epsilon_i \\
 \text{logit}(\xi_m) &= \delta_0 + u_m^{(\xi)} + \epsilon_i
 \end{aligned}$$

where $Y_{s,t,m}$ represents RX1day precipitation maximum for grid s , year t and dataset m . Following extreme value theory, $Y_{s,t,m}$ is modeled using the GEV distribution with $\mu_{s,t,m}$ (location), $\sigma_{s,t,m}$ (scale) and ξ_m (shape) parameters, that vary in space, time and dataset. The model is a HGAM (Pedersen et al., 2019) with intercepts ($\beta_0, \gamma_0, \delta_0$) and selected covariates accounting for long-term variability in time $\mathcal{F}(\text{year}(t))$, variability in space $\mathcal{G}(\text{lon}_s, \text{lat}_s)$ and dataset-specific deviations, $\mathcal{H}(\text{lon}_s, \text{lat}_s, u_m)$, that describe how each individual dataset m deviates from the overall spatial field $\mathcal{G}(\cdot)$. Functions $\mathcal{F}(\cdot)$, $\mathcal{G}(\cdot)$ and $\mathcal{H}(\cdot)$ are smooth functions of their respective covariates (year, longitude, latitude and dataset-specific random effects) that are estimated in the GAM fitting process. Within the GAM framework, these functions are a sum of smooth basis functions, so that for the $\mu_{s,t,m}$ parameter for example:

$$\begin{aligned}
 \mathcal{F}(\text{year}_t) &= \sum_{k=1}^K \beta_k b_k(\text{year}_t) \\
 \mathcal{G}(\text{lon}_s, \text{lat}_s) &= \sum_{m=1}^M \sum_{n=1}^N \beta_{m,n} c_n(\text{lat}_s) d_m(\text{lon}_s) \\
 \mathcal{H}(\text{lon}_s, \text{lat}_s, u_m) &= \sum_{p=1}^P \sum_{q=1}^Q \sum_{r=1}^R \beta_{p,q,r} e_r(u_m) \\
 &\quad y_q(\text{lat}_s) x_p(\text{lon}_s)
 \end{aligned}$$

where K, M and N are the number of knots that define the complexity (or wiggleness) of the basis functions $b(\cdot), c(\cdot), d(\cdot), x(\cdot)$ and $y(\cdot)$ with corresponding model coefficients β that are specific to each GEV parameter (but not denoted to avoid notational clutter). Equivalent functions for model coefficients γ and δ are also constructed. Note that $e(\cdot)$ is a special case whereby the dataset specific random effect is treated as a smooth function (see Wood, 2006, 2008; and Wood et al., 2013 for further details) where $e(\cdot)$ is a ridge spline basis that emulates an independent

and identically distributed Gaussian random effect. This adds as a constraint across data sets m for $\mathcal{H}(\cdot)$. The R code used to fit this model is available in the [Supplementary material](#).

Data blending is achieved via the “global” terms [the intercepts β_0, γ_0 and δ_0 , and functions $\mathcal{F}(\cdot)$ and $\mathcal{G}(\cdot)$] which are assumed common across m . Dataset-specific parameters $u_m^{(\cdot)}$ capture variability in the data (other than that explained by the global terms) due to discrepancies in each dataset. These parameters are assumed to be random effects and are modeled via:

$$\begin{aligned}
 u_m^{(\mu)} &\sim \text{Normal}(\mu_\mu, \sigma_\mu^2), \\
 u_m^{(\sigma)} &\sim \text{Normal}(\mu_\sigma, \sigma_\sigma^2), \\
 u_m^{(\xi)} &\sim \text{Normal}(0, \sigma_\xi^2).
 \end{aligned}$$

Where the parameters for these distributions are estimated from the model fitted parameters from $e(\cdot)$. Assuming these to be random quantities, explicitly captures variability from using multiple data sources, thus allowing predictions for datasets other than the ones used for estimation (e.g., radar measurements). These terms are designed to simulate within the model, additional unsampled dataset variability, as if we had incorporated more than m datasets. In fact, this formulation allows the uncertainty associated with different observed datasets to be integrated out for each GEV parameter e.g.,

$$\begin{aligned}
 \mu_{s,t} &= \int_{u_m^{(\mu)}} \mu_{s,t,m} du_m^{(\mu)} \\
 \sigma_{s,t} &= \int_{u_m^{(\sigma)}} \sigma_{s,t,m} du_m^{(\sigma)} \\
 \xi &= \int_{u_m^{(\xi)}} \xi_m du_m^{(\xi)}
 \end{aligned}$$

which are the blended estimates of the GEV parameters. Finally, predictions of the variable of interest are based on:

$$Y_{s,t} \sim \text{GEV}(\mu_{s,t}, \sigma_{s,t}, \xi).$$

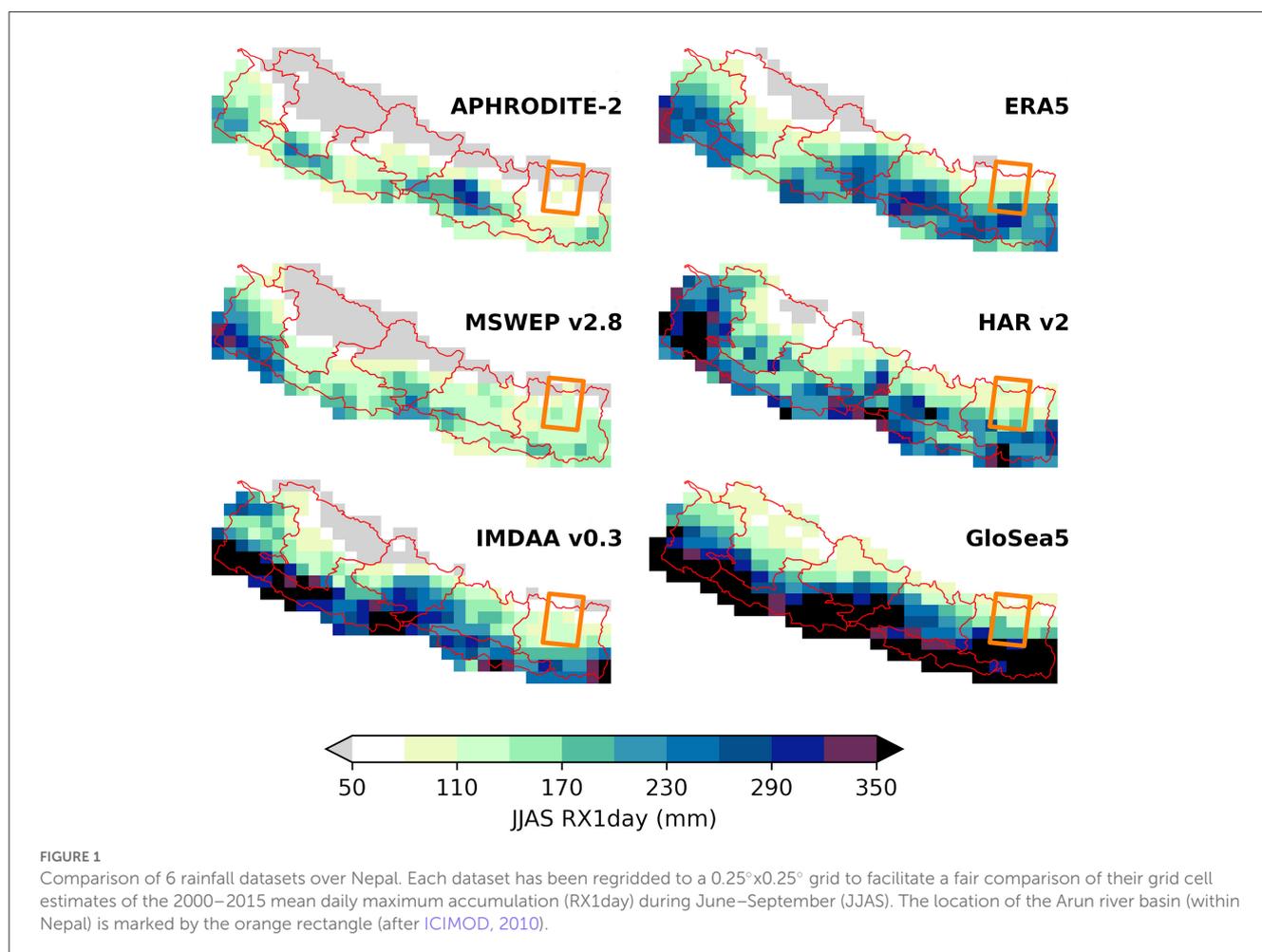
Although we illustrate the method using a GEV distribution fitted to rainfall extremes, the choice of predictive distribution, or mixture of distributions (e.g. Economou et al., 2023), should be tailored to the variable of interest. In some cases, suitable distributions may still produce unrealistic values. In this example, the GEV distribution may produce negative values for some GEV parameters, which would be unrealistic for rainfall. Identifying the most suitable predictive distribution may also be the primary limitation of this method. In some cases, it may be intractable to identify the mixture of distributions responsible for producing the data given conditional differences of the observing system.

3. Results

There are many sources of precipitation data over Nepal (e.g., Ceglar et al., 2017). For the purposes of demonstration, we examine five open-access datasets and one seasonal forecast system (summarized in Table 1), but in principle the method is agnostic to the number of datasets used. The differences in their

TABLE 1 Summary of observational datasets used in this study.

Name	Data source	URL access	References
Asian Precipitation—Highly-Resolved Observational Data Integration Toward Evaluation (APHRODITE-2) v1901	Rain gauges	http://aphrodite.st.hirosaki-u.ac.jp/index.html	Yatagai et al., 2012
Multi-Source Weighted-Ensemble Precipitation (MSWEP) v2.8	Merged rain gauge estimates with ERA5 and satellite data	http://www.gloh2o.org/mswep/	Beck et al., 2019
High Asia Refined analysis (HAR) v2	Dynamically downscaled reanalysis based on ERA5	https://www.klima.tu-berlin.de/index.php?show=daten_har2	Wang et al., 2021
Indian Monsoon Data Assimilation and Analysis (IMDAA) v0.3	Reanalysis	https://rds.ncmrwf.gov.in/	Rani et al., 2021
ERA5	Reanalysis	https://doi.org/10.24381/cds.adbb2d47	Hersbach et al., 2020
GloSea5	Ensemble seasonal prediction system	Not openly accessible	MacLachlan et al., 2015



estimates of the mean monsoon-total precipitation and monsoon maximum daily-total are presented in Figure 1. Each dataset has been conservatively regridded from their original resolution to a common 0.25°x0.25° grid, and timebound to the commonly shared period of 2000–2015. Even visually, the differences in spatial variability and magnitude of rainfall accumulation are apparent. From a policy makers’ perspective: should a new

hydropower plant in the Arun river basin (Koshi Pradesh, eastern Nepal, see Figure 1) plan to accommodate a maximum 1-day rainfall accumulation of <50–110 mm (APHRODITE-2) or >350 mm (GloSea5)?

The selected source datasets are representative of different methodological approaches to collecting and constructing gridded data. We choose ERA5, a global reanalysis dataset from

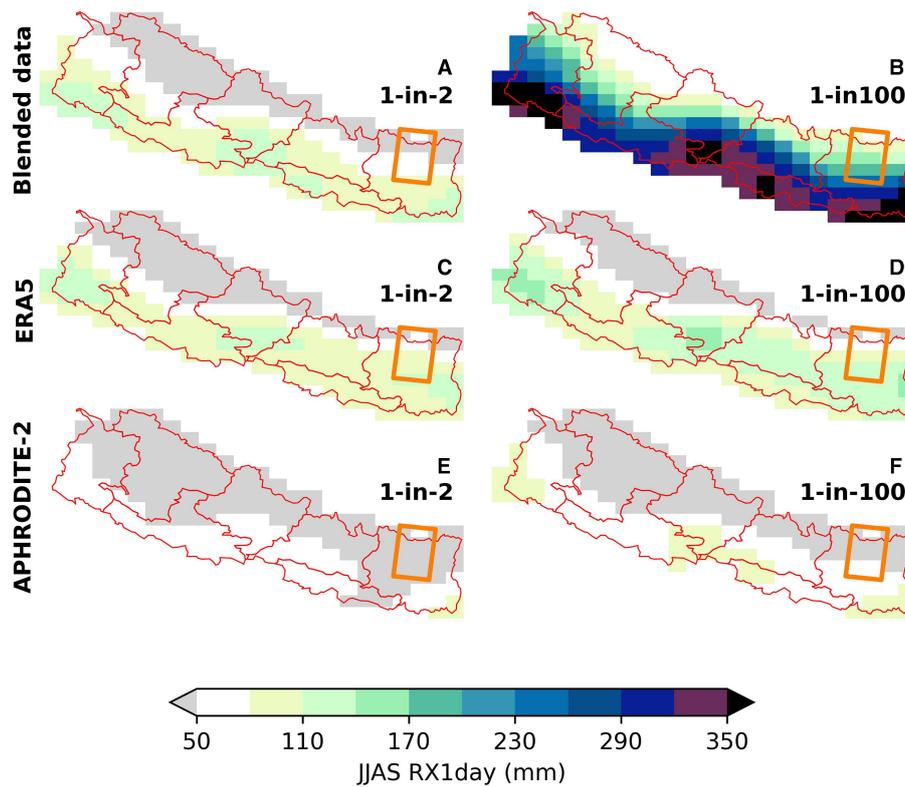


FIGURE 2
 Extreme value estimates of 1-in-2 and 1-in-100 year JJAS RX1day events over Nepal. (A, B) Extreme value estimates based on blended data estimate incorporating MSWEP, HAR, IMDAA and GloSea5 data, (C, D) ERA5 only and (E, F) APHRODITE-2 only.

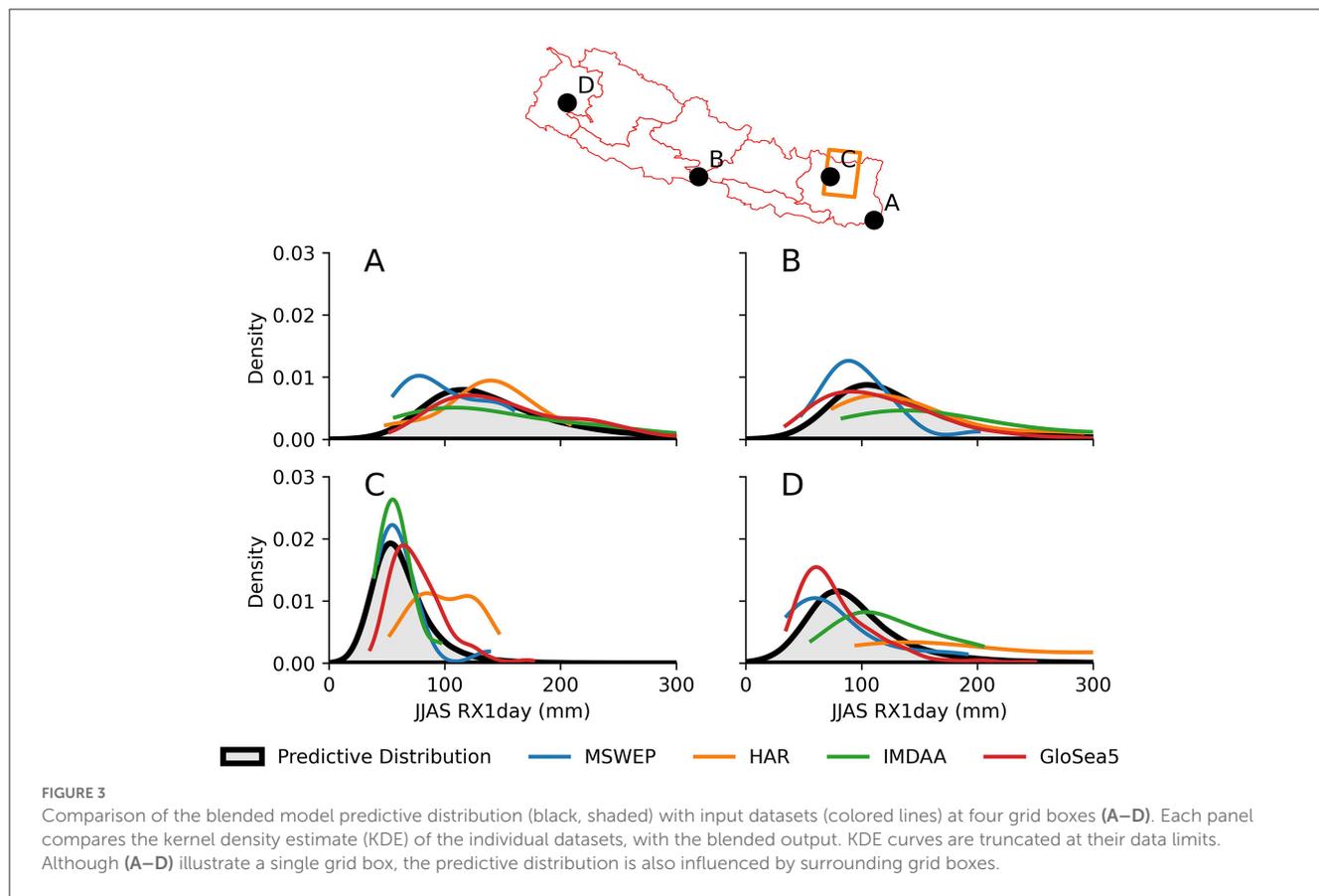
Hersbach et al. (2020), and APHRODITE-2, a gridded dataset based on *in-situ* rain gauge data from Yatagai et al. (2012), as baseline datasets against which we will compare our blending method. We apply our data blending framework to: MSWEP, a global combined dataset merging rain gauge estimates with ERA5 and satellite data from Beck et al. (2019); HAR v2 from Wang et al. (2021), a regional data set focusing on high mountain Asia, generated by dynamically downscaling ERA5 using the Weather Research and Forecasting model; IMDAA v0.3, a reanalysis dataset from Rani et al. (2021); and GloSea5 from MacLachlan et al. (2015), a seasonal prediction systems based on the HadGEM3 model. These four datasets produce a blended RX1day extreme value estimate. Figure 2 compares the blended result to the baseline datasets for estimates of 1-in-2 and 1-in-100 year RX1day events. RX1day estimates from a single dataset have the potential to significantly misrepresent rainfall accumulation, compared to an estimate derived from a greater number of data sources. This effect is more pronounced for extreme return periods.

We also illustrate the interpretability of this method by comparing the predictive distribution from the blended dataset, with the four input datasets (Figure 3). This comparison provides some insight into the data blending, whilst illustrating the consensus (or lack of) in the input datasets. For the blended dataset, the predictive distribution at each location illustrated in Figure 3 is the product of each of the four input data at that location (A–D), but is also influenced by the surrounding values as well.

For locations (A) and (C), there is significant variation between the four input datasets, such that the predictive distribution mediates the disparity, supporting the middle-ground. For (B), 3 of 4 datasets support a heavier tailed distribution than MSWEP, but the bulk of the predictive distribution is not as extreme as IMDAA. For (C), HAR is a clear outlier, but with strong agreement between MSWEP and IMDAA which has large influence on the predictive distribution.

4. Discussion

Often, ostensibly similar atmospheric datasets show marked differences in their estimate of reality. This is because their reliability is defined by the spatial coverage of surface stations, satellite algorithms, and the data assimilation models that contribute to their creation. Rain gauges provide relatively accurate and trusted measurements of precipitation at point locations, but are unavailable over many sparsely populated and oceanic areas. Satellite observations provide data with a greater degree of homogeneous spatial coverage, but contain non-negligible random errors and biases owing to the indirect nature of the relationship between the measurement from satellite-mounted instrumentation and precipitation at the Earth’s surface, inadequate temporal sampling given the satellite’s motion in space, and deficiencies in the data processing algorithms needed to amalgamate their



observations. Further datasets can be created by incorporating observations into numerical models that use mathematically defined physical processes to generate a synthesized estimate of precipitation across a uniform grid, with spatial homogeneity and temporal continuity.

Where a heuristic measure to identify a single best data source is impossible, and differences in data sources are large, it is difficult to a priori justify the use of a single data source in the decision making process. Approaches such as Bayesian melding (Poole and Raftery, 2000) have been proposed to this end, but with relatively little uptake, possibly due to the underlying complexity of the Bayesian framework, the lack of extension to non-Gaussian variables (such as precipitation) and the challenges associated with scaling this approach to large spatio-temporal datasets. Instead, we propose an approach based on Hierarchical Generalized Additive Models (HGAMs) (Hastie and Tibshirani, 1990; Wood, 2017; Pedersen et al., 2019). This flexible data modeling framework retains the ability to incorporate multiple sources of information and aggregate them into a single summary that is more informative than its constituent parts, but with easier model creation, good computational scalability and an ability to apply the method to non-Gaussian fields. The results of applying a HGAM to the data in Figure 1 are shown in Figure 2.

It is important that a data blending method retains an ability to discern how each separate dataset influences the blended output, such as in Figure 3 (i.e. it is interpretable), otherwise its utility for transparent decision making is no better than an arbitrary a priori judgment of an individual dataset. Uncertainty estimation is also a

key part of interpretability: to understand where there is agreement between data sources and appreciate the overall uncertainty of the resulting blended outcome. Crucially, the framework provided by GAMs allows the posterior predictive distribution to include all associated uncertainty (Wood, 2017). Probabilistic approaches have added value for decision making because they can be used in conjunction with decision theory to make rigorously repeatable decisions and quantifying the value (or utility) of the blended data.

Decision making requires quantification of risk and an appreciation of uncertainty. Whilst data only represents part of the decision-making landscape, data science for decision-making should focus on methods that provide interpretable and transparent frameworks for reproducible data blending. The proliferation of atmospheric datasets shouldn't be part of the decision makers' problem but, via data science, provide an approachable foundation on which to make informed decisions.

Data availability statement

The raw data supporting the conclusions of this article will be made available by the authors, without undue reservation.

Author contributions

Conceptualization, methodology, writing—original draft, and writing—review and editing: HS and TE. Investigation and

visualization: HS. All authors contributed to the article and approved the submitted version.

Funding

UK Aid as part of the Asia Regional Resilience to a Changing Climate (ARRCC) Programme. TE was supported by the European Union’s Horizon 2020 Research and Innovation Programme (EMME-CARE) under Grant Agreement No. 856612 https://ec.europa.eu/info/research-and-innovation/funding/funding-opportunities/funding-programmes-and-open-calls/horizon-europe_en and the Cyprus Government.

Acknowledgments

We thank two reviewers for their comments that helped us improve an initial version of this manuscript. ERA5 data is openly available from Copernicus Climate Change Service (C3S) (2023): ERA5 hourly data on single levels from 1940 to present. Copernicus Climate Change Service (C3S) Climate Data Store (CDS). doi: 10.24381/cds.adbb2d47. Our results contain modified Copernicus Climate Change Service information. Neither the European Commission nor ECMWF is responsible for any use that may be made of the Copernicus information or data it contains. The licence for Copernicus Products is available to view: <https://cds.climate.copernicus.eu/api/v2/terms/static/licence-to-use-copernicus-products.pdf>. For IMDAA, data is released under the Creative Commons Licence: Attribution-Non-Commercial-Share-Alike - 4.0 International (CC BY-NC-SA 4.0) and available from <https://rds.ncmrwf.gov.in/>. For HAR v2, data is openly available from <https://data.klima.tu-berlin.de/HAR/v2/>. For

MSWEP, data is released under the Creative Commons Attribution-NonCommercial 4.0 International (CC BY-NC 4.0) license and available from <http://www.gloh2o.org/mswep/>. GIS data relating to Nepal is available from the ICIMOD Regional Database System: <http://rds.icimod.org/Home/Index>. Model fitting was performed in R (R Core Team, 2022) using the mgcv package (Wood, 2017). Plotting was done in python using matplotlib (Hunter, 2007; Caswell et al., 2021) and iris (Met Office, 2013; Hattersley et al., 2021).

Conflict of interest

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.

Publisher’s note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Supplementary material

The Supplementary Material for this article can be found online at: <https://www.frontiersin.org/articles/10.3389/fdata.2023.1198097/full#supplementary-material>

References

Basnyat, D. B., and Watkiss, P. (2017). *Adaptation to Climate Change in the Hydroelectricity Sector in Nepal*. Climate and Development Knowledge Network. Available online at: <https://cdkn.org/sites/default/files/files/hydroelectricity-in-Nepal-final-WEB.pdf> (accessed January 24, 2023).

Beck, H. E., Wood, E. F., Pan, M., Fisher, C. K., Miralles, D. G., van Dijk, A. I. J. M., et al. (2019). MSWEP V2 Global 3-Hourly 0.1° precipitation: methodology and quantitative assessment. *Bull. Am. Meteorol. Soc.* 100, 473–500. doi: 10.1175/BAMS-D-17-0138.1

Caswell, T. A., Droettboom, M., Lee, A., de Andrade, E. S., Hoffmann, T., Hunter, J., et al. (2021). *matplotlib/matplotlib: REL: v3.5.1*. Honolulu, HI: Zenodo.

Ceglar, A., Toreti, A., Balsamo, G., and Kobayashi, S. (2017). Precipitation over monsoon asia: a comparison of reanalyses and observations. *J. Clim.* 30, 465–476. doi: 10.1175/JCLI-D-16-0227.1

DHM (2022). *Normal Climate Data Portal*. Available online at: <https://www.dhm.gov.np/climate-services/normal-climate> (accessed September 30, 2022).

Economou, T., Lazoglou, G., Tzyrkalli, A., Constantinidou, K., and Lielieveld, J. (2023). A data integration framework for spatial interpolation of temperature observations using climate model data. *PeerJ* 11, e14519. doi: 10.7717/peerj.14519

Hastie, T. J., and Tibshirani, R. J. (1990). *Generalized Additive Models*. 1st ed. New York, NY: Routledge.

Hattersley, R., Little, B., Elson, P., Campbell, E., Peglar, P., Killick, P., et al. (2021). *SciTools/iris: v3.0.3*. Honolulu, HI: Zenodo.

Hersbach, H., Bell, B., Berrisford, P., Hirahara, S., Horányi, A., Muñoz-Sabater, J., et al. (2020). The ERA5 global reanalysis. *Q. J. R. Meteorol. Soc.* 146, 1999–2049. doi: 10.1002/qj.3803

Hunter, J. D. (2007). Matplotlib: A 2D graphics environment. *Comp. Sci. Eng.* 9, 90–95. doi: 10.1109/MCSE.2007.55

ICIMOD (2010). *Sub-sub-basins of Hindu Kush Himalaya (HKH) Region*. Kathmandu: ICIMOD.

MacLachlan, C., Arribas, A., Peterson, K. A., Maidens, A., Fereday, D., Scaife, A. A., et al. (2015). Global seasonal forecast system version 5 (GloSea5): a high-resolution seasonal forecast system. *Q. J. R. Meteorol. Soc.* 141, 1072–1084. doi: 10.1002/qj.2396

McCright, A. M., and Dunlap, R. E. (2000). Challenging global warming as a social problem: an analysis of the conservative movement’s counter-claims. *Soc. Prob.* 47, 499–522. doi: 10.2307/3097132

Met Office (2013). *Iris: A Python Package for Analysing and Visualising Meteorological and Oceanographic Data Sets*. Exeter, Devon. Available online at: <http://scitools.org.uk/> (accessed January 24, 2023).

Orlove, B., Shwom, R., Markowitz, E., and Cheong, S.-M. (2020). Climate decision-making. *Ann. Rev. Environ. Resou.* 45, 271–303. doi: 10.1146/annurev-environ-012320-085130

- Pedersen, E. J., Miller, D. L., Simpson, G. L., and Ross, N. (2019). Hierarchical generalized additive models in ecology: an introduction with mgcv. *PeerJ* 7, e6876. doi: 10.7717/peerj.6876
- Poole, D., and Raftery, A. E. (2000). Inference for deterministic simulation models: the bayesian melding approach. *J. Am. Stat. Assoc.* 95, 1244–1255. doi: 10.1080/01621459.2000.10474324
- R Core Team (2022). *R: A Language and Environment for Statistical Computing*. Vienna, Austria. Available online at: <https://www.r-project.org> (accessed January 24, 2023).
- Rani, S. I., Arulalan, T., George, J. P., Rajagopal, E. N., Renshaw, R., Maycock, A., et al. (2021). IMDAA: high resolution satellite-era reanalysis for the indian monsoon region. *J. Clim.* 34, 5109–5133. doi: 10.1175/JCLI-D-20-0412.1
- Wang, X., Tolksdorf, V., Otto, M., and Scherer, D. (2021). WRF-based dynamical downscaling of ERA5 reanalysis data for high mountain Asia: towards a new version of the high asia refined analysis. *Int. J. Climatol.* 41, 743–762. doi: 10.1002/joc.6686
- Wood, S. N. (2006). Low-rank scale-invariant tensor product smooths for generalized additive mixed models. *Biometrics* 62, 1025–1036. doi: 10.1111/j.1541-0420.2006.00574.x
- Wood, S. N. (2008). Fast stable direct fitting and smoothness selection for generalized additive models. *J. R. Stat. Soc. B Stat. Methodol.* 70, 495–518. doi: 10.1111/j.1467-9868.2007.00646.x
- Wood, S. N. (2017). *Generalized Additive Models: An Introduction with R*. 2nd ed. London: Chapman and Hall/CRC.
- Wood, S. N., Scheipl, F., and Faraway, J. J. (2013). Straightforward intermediate rank tensor product smoothing in mixed models. *Stat. Comp.* 23, 341–360. doi: 10.1007/s11222-012-9314-z
- Yatagai, A., Kamiguchi, K., Arakawa, O., Hamada, A., Yasutomi, N., and Kitoh, A. (2012). APHRODITE: constructing a long-term daily gridded precipitation dataset for Asia based on a dense network of rain gauges. *Bull. Am. Meteorol. Soc.* 93, 1401–1415. doi: 10.1175/BAMS-D-11-00122.1