frontiers

Supplementary Material

Hybrid Modeling of Drop Breakage in Pulsed Sieve Tray Extraction Columns

Andreas Palmtag[1], Johannes Rousselli[1], Henning Gröschl[1], Andreas Jupke[1*]

1) Fluid Process Engineering (AVT.FVT), RWTH Aachen University, Forckenbeckstraße 51, D-52074 Aachen, Germany, Tel.: +49 241 80-95490, www.avt.rwth-aachen.de, e-mail: andreas.jupke@avt.rwth-aachen.de

*** Correspondence:**
Andreas Jupke
andreas.jupke@avt.rwth-aachen.de

1 Supplementary Tables

Table 1: Parametrization of Garthe's breakage model.

Solvent system	d_h [mm]	c_1	c_2	c_3	c_4
TW	2	1.64	-0.18	1..91	0.55
	4	4.8	0.27	1.35	4.31
TWA	2	3.81	0.61	1.11	3.47
	4	4.75	0.14	1.11	4.35
BW	2	1.33	0.03	2.03	0.42
	4	2.00	-0.07	1.61	0.95
BWA	2	2.49	0.27	0.95	1.77
	4	2.18	-0.33	1.62	1.15

2 Error Metric

The root mean-squared error e_{rmse} is commonly used when assessing the deviation between a number n_{u} of predicted u and experimental values $\hat{u}$ (Brockkötter et al., 2020). The e_{rmse} is averaged over all n_{u} data sets, thus it might be disproportionally affected by outliers (Dahmen & Reusken, 2022). Nevertheless, e_{rmse} is a common measure of the residue in machine learning (ML) since it provides the information on the residue in its most simple form, e.g., a small e_{rmse} indicates a good model, a large e_{rmse} quantifies the average deviation in the same dimension as the predicted quantity. Commonly, a distinction is made between the training e_{rmse}, the value minimized during model development, and test e_{rmse}, a score assessing the prediction of the ML model on data not used during model development (James & Witten, 2013).

In contrast to e_{rmse}, the coefficient of determination e_{R2} is a relative measure of the residue. e_{R2} is common in regression analysis, and it can be interpreted as a comparison between the deviation $\hat{u} - u$ to the average of the experimental values $\bar{u}$. At best, the e_{R2} is close to 1, whereas, a $e_{\mathrm{R2}} < 0$ indicates that the experimental values are better represented by $\bar{u}$ than by the predictions u. (Cramer & Kamps, 2017; James & Witten, 2013)

The pull metric e_{pull} is not commonly used in the extraction research. Therefore, we would like to demonstrate the pull metric based on a simple example. We consider a database consisting of $n_{\mathrm{u}} = 100$ experimental values $\hat{u}$ and the according predictions u by a model. The n_{u} experimental values represent independent experimental data sets and not replicates of one experiment. For each of the n_{u} entries in the database, e.g., measurement-predictions pairs, the deviation $\hat{u} - u$ is calculated and standardized by the measurement uncertainty σ_e, yielding the e_{pull} for each entry (compare with eq. 4-3). The resulting n_{u} pull values e_{pull} represent a population which can be visualized in a histogram (see Figure 1). The resulting distribution is characterized by the mean $\bar{e}_{\mathrm{pull}}$ and its standard deviation $\tilde{e}_{\mathrm{pull}}$. Considering the numerical values, a good model is characterized by a pull distribution with a mean close to zero ($\bar{e}_{\mathrm{pull}} = 0$) and a standard deviation smaller than one ($\tilde{e}_{\mathrm{pull}} < 1$). Graphically, a good distribution has its center close $\bar{e}_{\mathrm{pull}} = 0$ and most entries within the $-1 \leq e_{\mathrm{pull}} \leq 1$ (indicated by dashed lines in Figure 1), indicating that most entries in the database have a deviation that does not exceed the measurement uncertainty. It is important to note that the numerical values for $\bar{e}_{\mathrm{pull}}$ and $\tilde{e}_{\mathrm{pull}}$ might not suffice to assess the accuracy of the prediction, since a multimodal distribution might also result in allegedly good values for $\bar{e}_{\mathrm{pull}}$ and $\tilde{e}_{\mathrm{pull}}$. Therefore, we also considered the graphical representation of the pull distribution to assess the prediction quality of our models.

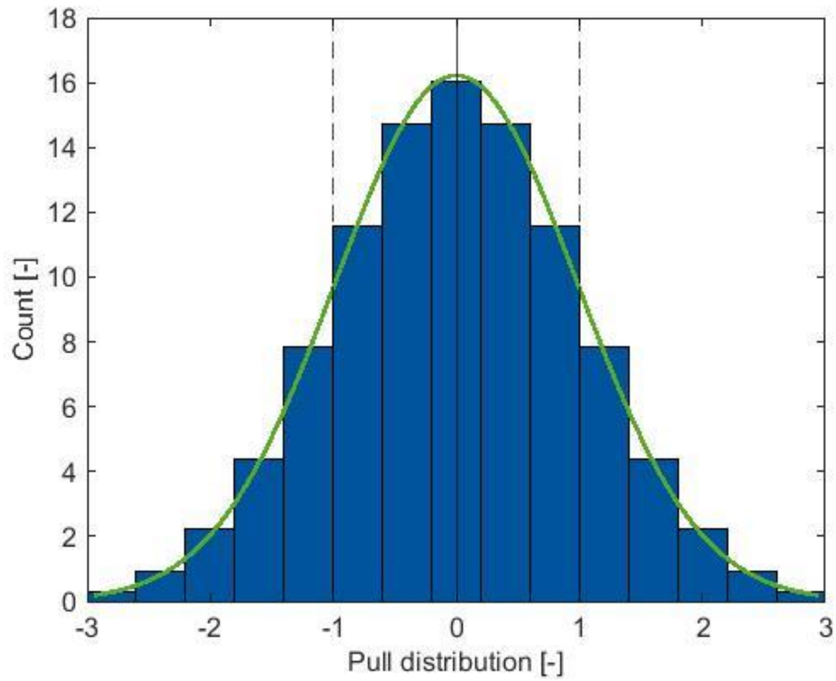

Figure 1: Exemplary Pull distribution.

3 Specification of the Soft- and Hardware

All simulations have been conducted on a desktop computer with an Intel Core i7-7700K @4.2GHz processor, which has 4 cores and 8 GB of RAM. On overview over the python and Matlab libraries is given in Table 2.

Table 2: Specification of the software used in this work.

Name	Type	Version
Python	Programming Language	3.10.8. 64-bit
numpy	Python library	1.23.4
scipy	Python library	1.9.3
pandas	Python library	1.5.1
XlsxWriter	Python library	3.0.3
joblib	Python library	1.2.0
scikit-learn	Python library	1.1.3
openpyxl	Python library	3.0.10
torch	Python library	1.13.0
colorama	Python library	0.4.6
fluids	Python library	1.0.22
mlxtend	Python library	0.21.0
seaborn	Python library	0.12.1
Matlab™	Software	R2022b
Optimization Toolbox	Library (Matlab™)	9.4
Curve Fitting Toolbox	Library (Matlab™)	3.8
Parallel Computing Toolbox	Library (Matlab™)	7.7
Deep Learning Toolbox	Library (Matlab™)	14.5
Statistics & Machine Learning Toolbox	Library (Matlab™)	12.4.

4 References

Brockkötter, J., Cielanga, M., Weber, B., & Jupke, A. (2020). Prediction and Characterization of Flooding in Pulsed Sieve Plate Extraction Columns Using Data-Driven Models. *Industrial & Engineering Chemistry Research*, *59*(44), 19726–19735. https://doi.org/10.1021/acs.iecr.0c03282

Cramer, E., & Kamps, U. (2017). *Grundlagen der Wahrscheinlichkeitsrechnung und Statistik: Eine Einführung für Studierende der Informatik, der Ingenieur- und Wirtschaftswissenschaften* (4., korrigierte und erweiterte Auflage). *Springer-Lehrbuch*. Springer Spektrum.

Dahmen, W., & Reusken, A. (2022). *Numerik für Ingenieure und Naturwissenschaftler* (3. Auflage). *Springer-Lehrbuch*. Springer. https://doi.org/10.1007/978-3-540-76493-9

James, G., & Witten, D. (2013). *#x88;An introduction to statistical learning: with applications in R.*