

metrica: an R package to evaluate prediction performance of regression and classification point-forecast models

Adrian A. Correndo^{1¶}, Luiz H. Moro Rosso², Carlos H. Hernandez¹, Leonardo M. Bastos⁵, Luciana Nieto¹, Dean Holzworth⁴, and Ignacio A. Ciampitti¹

¹ Department of Agronomy, Kansas State University, Manhattan, KS, USA. ² Private Consultant, Brasil. ³ Department of Crop and Soil Sciences, University of Georgia, Athens, GA, USA. ⁴ CSIRO Agriculture and Food, Australia. [¶] Corresponding author

DOI: [10.xxxxxx/draft](https://doi.org/10.xxxxxx/draft)

Software

- [Review](#)
- [Repository](#)
- [Archive](#)

Editor: [Open Journals](#)

Reviewers:

- [@openjournals](#)

Submitted: 01 January 1970

Published: unpublished

License

Authors of papers retain copyright and release the work under a Creative Commons Attribution 4.0 International License ([CC BY 4.0](#)).

Summary



The *metrica* R package (Correndo et al., 2022) is an open-source software designed to facilitate the quantitative and visual assessment of prediction performance of point-forecast simulation models for continuous (regression) and categorical variables (classification). The package ensembles a series of 80+ functions that account for multiple aspects of the agreement between predicted and observed values. Without the need of advanced skills on programming, *metrica* enables users to automate the estimation of multiple prediction performance metrics including goodness of fit, error metrics, error decomposition, model efficiency, indices of agreement, and to produce stylish data visualization outputs. This article introduces *metrica*, an R package developed with the main objective of contributing to transparent and reproducible evaluation of point-forecast models performance.

Statement of need

Evaluating the prediction quality is a crucial step for any simulation model, for which a myriad of metrics and visualization techniques have been developed (Tedeschi, 2006; Wallach et al., 2019; Yang et al., 2014). Nonetheless, to conduct a comprehensive assessment of the predicted-observed agreement in R (R Core Team, 2021), users normally have to rely on multiple packages, and even on self-defined functions, which increases the risk of involuntary mistakes due to the need of fluctuating syntax and data wrangling.

As the reproducibility of data analysis continues to be a challenge for science (Seibold, 2022), developing open source software like *metrica* offers a step toward a transparent and reproducible process to assist researchers in evaluating models performance. We decided to create *metrica* in R (R Core Team, 2021) due to its substantial role in data science (Thieme, 2018). Under its open-source philosophy, R empowers the democratization of statistical

33 computing ([Hackenberger, 2020](#)) by hosting and globally distributing cutting-edge algorithms
34 through the Comprehensive R Archive Network (CRAN).

35 Finally, it is noteworthy that in the area of agricultural sciences, although point-forecast simu-
36 lation models such as the Agricultural Production Systems sIMulator (APSIM) ([D. Holzworth
37 et al., 2018](#); [D. P. Holzworth et al., 2014](#)) count with tools to facilitate the integration into R
38 through packages such as *apsimx* ([Miguez, 2022](#)), the assessment of its prediction quality is not
39 yet integrated for R users. Therefore, we aim for *metrica* to offer users of simulation models
40 for agriculture, plant, and soil sciences community a toolbox for assessing the performance of
41 regression and classification point-forecast models.

42 Package features

43 For regression models, *metrica* includes four plotting functions (scatter, tiles, density, &
44 Bland-Altman plots) using *ggplot2* ([Wickham, 2016](#)), and 48 prediction performance metrics.
45 For classification models (two-class or multi-class), it includes one function to visualize a
46 confusion matrix, and 27 functions of prediction scores. The full list of metrics with description,
47 formula, and literature sources is presented in the package documentation at:

- 48 ■ Regression metrics: [https://adriancorrendo.github.io/metrica/articles/available_met-](https://adriancorrendo.github.io/metrica/articles/available_metrics_regression.html)
49 [rics_regression.html](https://adriancorrendo.github.io/metrica/articles/available_metrics_regression.html).
- 50 ■ Classification metrics: [https://adriancorrendo.github.io/metrica/articles/available_met-](https://adriancorrendo.github.io/metrica/articles/available_metrics_classification.html)
51 [rics_classification.html](https://adriancorrendo.github.io/metrica/articles/available_metrics_classification.html).

52 To extent of our knowledge, *metrica* covers several functions not supported, or partially sup-
53 ported by similar R packages (or components) designed for model evaluation such as *yardstick*
54 ([Kuhn & Vaughan, 2022](#)) from *tidymodels* ([Kuhn & Wickham, 2020](#)), the measuring perfor-
55 mance components from *caret* ([Kuhn, 2022](#)) or *mlr3* ([Lang et al., 2019](#)), *Metrics* ([Hamner
56 & Frasco, 2018](#)), *hydroGOF* ([Zambrano-Bigiarini, 2020](#)), *cvms* ([Olsen & Zachariae, 2021](#)),
57 *scoringutils* ([Bosse et al., 2020](#)), or *performance* ([Lüdecke et al., 2021](#)). Unique features
58 include:

- 59 ■ the most extensive collection of prediction performance metrics for regression and
60 classification models up to date.
- 61 ■ working under both vectorized (calling variables with `$`) or tabulated forms ([Wickham et
62 al., 2019](#)).
- 63 ■ controlling the output format as a list (`tidy = FALSE`) or as a table (`tidy = TRUE`).
- 64 ■ for classification, functions automatically recognizing two-class or multi-class data; and
65 specifically for multi-class cases, several metrics can be estimated for each class (`atom
66 = TRUE`) ([Ferri et al., 2009](#)), ([Ben-David, 2007](#)), including balanced and imbalanced
67 scenarios ([Kubat et al., 1997](#)).
- 68 ■ for regression, implementing a symmetric linear regression (standardized major axis-SMA-,
69 ([Warton et al., 2006](#))) to describe: i) pattern of the bivariate relationship with linear
70 parameters (`B0_sma`, `B1_sma`), and ii) degree of predicted-observed agreement by using
71 SMA-line to decompose the mean-squared-error (MSE) into lack of accuracy (MLA, PLA,
72 RMLA) and lack of precision (MLP, PLP, RMLP) components ([Correndo et al., 2021](#)).
- 73 ■ offering MSE decomposition approaches described by ([Kobayashi & Salam, 2000](#)) (SB,
74 SDS, LCS), and ([Smith & Rose, 1995](#)) (`Ub`, `Uc`, `Ue`).
- 75 ■ including multiple indices of agreement and model efficiency such as: i) index of agreement
76 `d` ([Willmott, 1981](#)), and its modified `d1` ([Willmott et al., 1985](#)) and refined `d1r` ([Willmott
77 et al., 2012](#)) variants, ii) Nash–Sutcliffe model efficiency (NSE) ([Nash & Sutcliffe, 1970](#))
78 and its improved variants `E1` ([Legates & McCabe Jr., 1999](#)), `Erel` ([Krause et al., 2005](#)),

79 and Kling-Gupta model efficiency (KGE) (Kling et al., 2012), iii) Robinson's index of
80 agreement (RAC) (Robinson, 1957, 1959), iv) Ji & Gallo agreement coefficient (AC) (Ji
81 & Gallo, 2006), v) Duvellier's Lambda (Duvellier & Meroni, 2016), vi) distance correlation
82 (dcorr) (Székely et al., 2007), or vii) maximal information coefficient (MIC) (Reshef et
83 al., 2011)), among others.

- 84 ■ importing files from APSIM Classic with `import_apsim_out()`, and from APSIM Next
85 Generation with the `import_apsim_db()` function.

86 System requirements and installation

87 Since *metrica* operates within R, the first step is to install R ($\geq 4.2.0$). To install and load
88 the package:

```
# Stable version (CRAN)
install.packages("metrica")

# Development version (GitHub)
devtools::install_github("adriancorrendo/metrica")

# Load
library(metrica)
```

89 Using the functions

90 There are two core arguments to all *metrica* functions: (i) `obs` (O_i ; observed, a.k.a. actual,
91 measured, truth, target, label), and (ii) `pred` (P_i ; predicted, a.k.a. simulated, fitted, modeled,
92 estimate) values. For regression, specific functions require defining the axis orientation (e.g.
93 predicted vs. observed -PO- or observed vs. predicted -OP-).

94 For two-class models, the `pos_level` argument serves to indicate the alphanumeric order of the
95 "positive level". Following most two-class denominations as `c(0,1)`, `c("Negative", "Positive")`,
96 and `c("FALSE", "TRUE")`, the default `pos_level = 2` (1, "Positive", "TRUE"). However,
97 we recognize other cases as possible (e.g. `c("Crop", "NoCrop")`), for which the user needs
98 to specify `pos_level = 1`. For multi-class classification, some functions present the `atom`
99 argument (TRUE / FALSE), which controls the output to be an overall average estimate
100 across all classes (default), or class-wise.

101 Example 1: Regression (continuous variables)

102 The following lines of code serve to run basic regression performance analysis using a native
103 dataset called `wheat`.

```
# Define dataset
data_wheat <- metrica::wheat

# Estimate Root Mean Square Error, result as a list
RMSE(data = data_wheat, obs = obs, pred = pred, tidy = FALSE)
#> $RMSE
#> [1] 1.666441

# Store results as a data frame
RMSE(data = data_wheat, obs = obs, pred = pred, tidy = TRUE)
```

104 To estimate multiple regression metrics at once using `metrica::metrics_summary()`:

```
# Define metrics list
my_reg_metrics <- c("R2", "CCC", "MBE", "RMSE", "RSR", "NSE", "KGE")
```

```
# Run metrics summary
metrics_summary(data = data_wheat,
  obs = obs, pred = pred,
  type = "regression",
  metrics_list = my_reg_metrics)
```

```
#>      Metric      Score
#> 1      R2      0.84555376
#> 2      CCC      0.91553253
#> 3      MBE      0.31815953
#> 4      RMSE      1.66644142
#> 5      RRMSE      0.19094834
#> 6      RSR      0.09678632
#> 7      PLP      5.15949064
#> 8      PLA      94.84050936
#> 9      NSE      0.83871126
#> 10     KGE      0.91064709
```

105 To produce a scatter plot of predicted vs. observed values as a customizable ggplot object:

```
scatter_plot(data = data_wheat,
  obs = obs , pred = pred,
  orientation = "P0",
  print_metrics = TRUE,
  metrics_list = my_reg_metrics,
  print_eq = TRUE,
  position_eq = c(x=14, y = 2),
  # Optional arguments to customize the plot
  shape_type = 21,
  shape_color = "steelblue",
  shape_size = 3,
  regline_type = "F1",
  regline_color = "#9e0059",
  regline_size = 2)+
# Customize axis breaks
scale_y_continuous(breaks = seq(0,20, by = 2))+
scale_x_continuous(breaks = seq(0,20, by = 2))
```

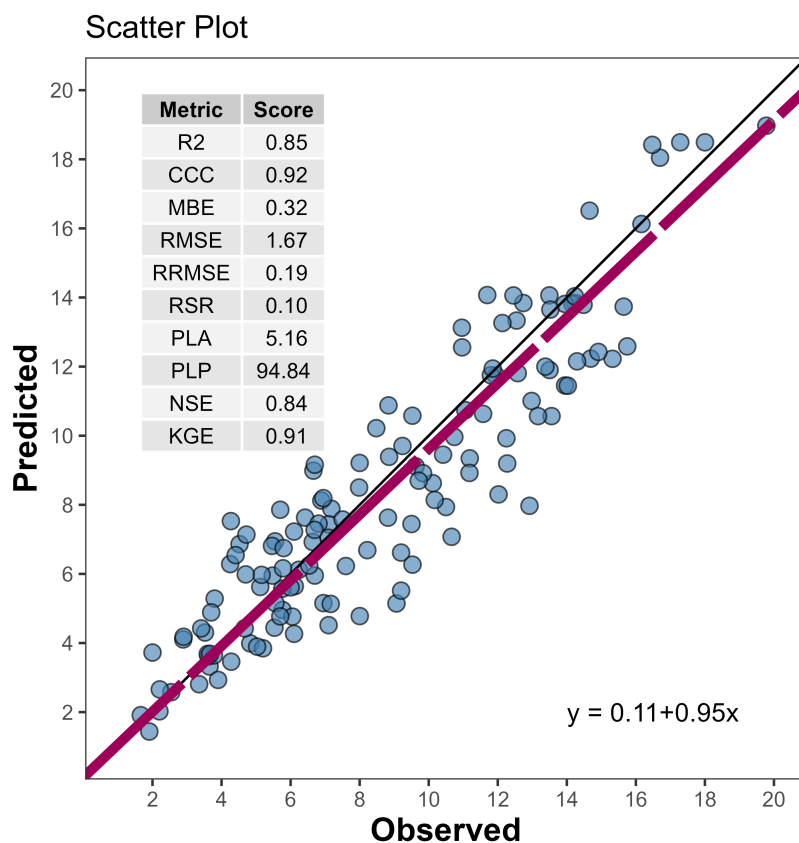


Figure 1: Predicted vs. Observed scatter plot using `metrica::scatter_plot()`.

106 Example 2: Classification (categorical variables)

107 The following lines of code serve to run a basic classification performance analysis using a
108 native dataset called `maize_phenology`.

```
# Define dataset
data_multiclass <- metrica::maize_phenology

# Estimate accuracy, result as a list
accuracy(data = data_multiclass, obs = actual, pred = predicted, tidy = FALSE)

#> $accuracy
#> [1] 0.8834951

# Result as a data frame
accuracy(data = data_multiclass, obs = actual, pred = predicted, tidy = TRUE)

#>      Metric      Score
#> 1    accuracy 0.8834951

# Define selected metrics
my_class_metrics <- c("accuracy", "precision", "recall", "specificity",
                      "fscore", "gmean", "khat")

# Run the summary for selected metrics
metrics_summary(data = data_multiclass,
```

```
obs = actual, pred = predicted,  
type = "classification")
```

```
#>      Metric      Score  
#> 1  accuracy 8.834951e-01  
#> 2  precision 8.335108e-01  
#> 3   recall 8.405168e-01  
#> 4 specificity 9.915764e-01  
#> 5   fscore 8.369991e-01  
#> 6    agf 8.370017e-01  
#> 7   gmean 9.129275e-01  
#> 8    khat 8.624527e-01
```

109 To produce a confusion matrix plot users may use:

```
confusion_matrix(data = data_multiclass,  
  obs = actual, pred = predicted,  
  plot = TRUE,  
  colors = c(low="grey85" , high="steelblue"),  
  unit = "count",  
  # Print metrics_summary  
  print_metrics = TRUE,  
  # List of performance metrics  
  metrics_list = my_class_metrics,  
  # Position (bottom or top)  
  position_metrics = "bottom")
```

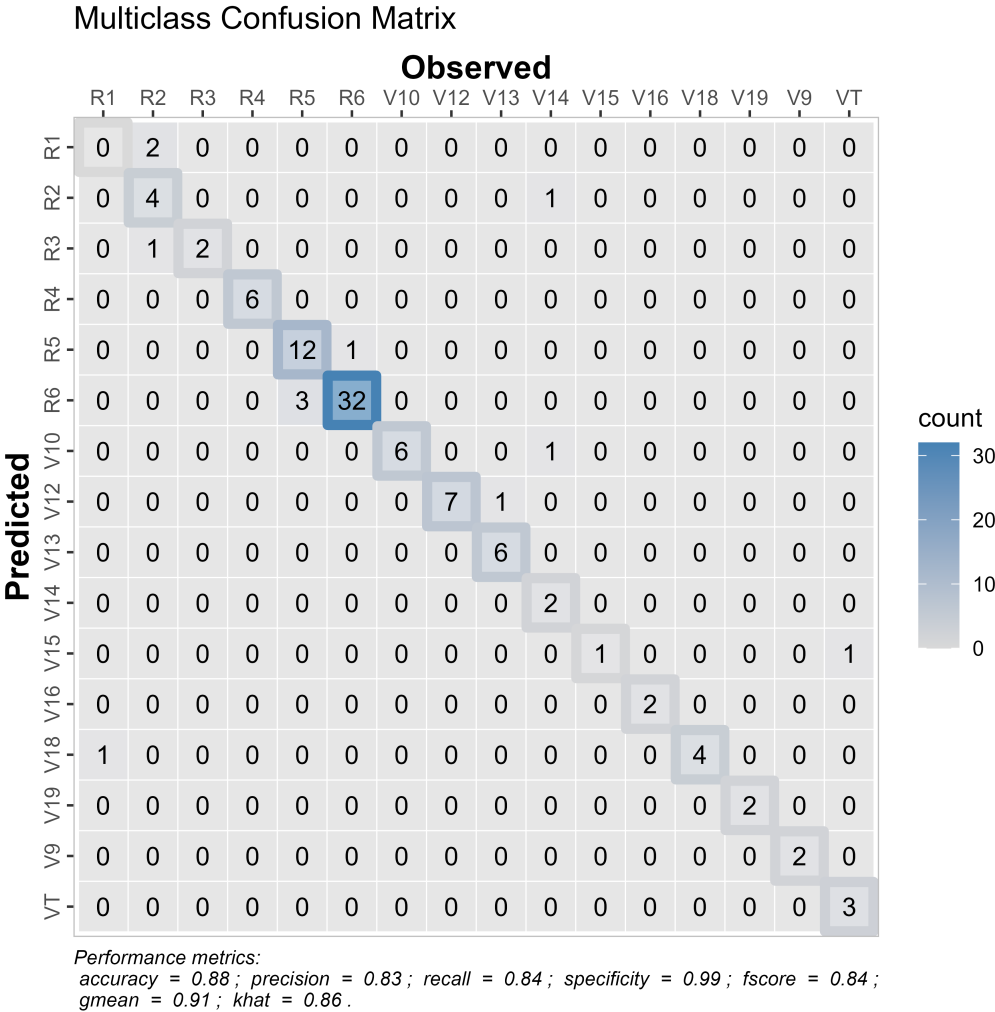


Figure 2: Confusion matrix plot using `metrca::confusion_matrix()`.

Documentation & License

The complete documentation and vignettes of the package can be found online at <https://adriancorrendo.github.io/metrca/>. *metrca* is under the MIT License (<https://opensource.org/licenses/MIT>). Source code is available at GitHub (<https://github.com/adriancorrendo/metrca>) along with its corresponding section to report issues and suggestions (<https://github.com/adriancorrendo/metrca/issues>).

Acknowledgements

Authors gratefully acknowledge the financial support by the Feed the Future Innovation Lab for Collaborative Research on Sustainable Intensification (SIIL) at Kansas State University through funding United States Agency for International Development (USAID) under the Cooperative Agreement (Grant number AID-OAA-L-14-00006).

References

- Ben-David, A. (2007). A lot of randomness is hiding in accuracy. *Engineering Applications of Artificial Intelligence*, 20, 875–885. <https://doi.org/10.1016/j.engappai.2007.01.001>
- Bosse, N. I., Gruson, H., Funk, S., EpiForecasts, & Abbott, S. (2020). *Scoringutils: Utilities for scoring and assessing predictions*. <https://doi.org/10.5281/zenodo.4618017>
- Correndo, A. A., Hefley, T. J., Holzworth, D. P., & Ciampitti, I. A. (2021). Revisiting linear regression to test agreement in continuous predicted-observed datasets. *Agricultural Systems*, 192, 103194. <https://doi.org/10.1016/j.agsy.2021.103194>
- Correndo, A. A., Moro Rosso, L. H., Schwalbert, R., Hernandez, C., Bastos, L. M., Nieto, L., Holzworth, D., & Ciampitti, I. A. (2022). *Metrica: Prediction performance metrics*. <https://CRAN.R-project.org/package=metrica>
- Duveiller, F., G., & Meroni, M. (2016). Revisiting the concept of a symmetric index of agreement for continuous datasets. *Scientific Reports*, 6, 19401. <https://doi.org/10.1038/srep19401>
- Ferri, C., Hernández-Orallo, J., & Modroiu, R. (2009). An experimental comparison of performance measures for classification. *Pattern Recognition Letters*, 30, 27–38. <https://doi.org/10.1016/j.patrec.2008.08.010>
- Hackenberger, B. K. (2020). R software: Unfriendly but probably the best. *Croat Med. J.*, 29;61(1), 66–68. <https://doi.org/10.3325/cmj.2020.61.66>
- Hamner, B., & Frasco, M. (2018). *Metrics: Evaluation metrics for machine learning*. <https://CRAN.R-project.org/package=Metrics>
- Holzworth, D. P., Huth, N. I., deVoil, P. G., Zurcher, E. J., Herrmann, N. I., McLean, G., Chenu, K., van Oosterom, E. J., Snow, V., Murphy, C., Moore, A. D., Brown, H., Whish, J. P. M., Verrall, S., Fainges, J., Bell, L. W., Peake, A. S., Poulton, P. L., Hochman, Z., ... Keating, B. A. (2014). APSIM – evolution towards a new generation of agricultural systems simulation. *Environmental Modelling & Software*, 62, 327–350. <https://doi.org/10.1016/j.envsoft.2014.07.009>
- Holzworth, D., Huth, N. I., Fainges, J., Brown, H., Zurcher, E., Cichota, R., Verrall, S., Herrmann, N. I., Zheng, B., & Snow, V. (2018). APSIM next generation: Overcoming challenges in modernising a farming systems model. *Environmental Modelling & Software*, 103, 43–51. <https://doi.org/10.1016/j.envsoft.2018.02.002>
- Ji, L., & Gallo, K. (2006). An agreement coefficient for image comparison. *Photogrammetric Engineering & Remote Sensing*, 72(7), 823–833. <https://doi.org/10.14358/PERS.72.7.823>
- Kling, H., Fuchs, M., & Paulin, M. (2012). Runoff conditions in the upper danube basin under an ensemble of climate change scenarios. *Journal of Hydrology*, 424–425, 264–277. <https://doi.org/10.1016/j.jhydrol.2012.01.011>
- Kobayashi, K., & Salam, M. U. (2000). Comparing simulated and measured values using mean squared deviation and its components. *Agronomy Journal*, 92(2), 345–352. <https://doi.org/10.2134/agronj2000.922345x>
- Krause, P., Boyle, D. P., & Bäse, F. (2005). Comparison of different efficiency criteria for hydrological model assessment. *Advances in Geosciences*, 5, 89–97. <https://doi.org/10.5194/adgeo-5-89-2005>
- Kubat, M., Matwin, S., & others. (1997). Addressing the curse of imbalanced training sets: One-sided selection. *ICML*, 97, 179.

- 166 Kuhn, M. (2022). *Caret: Classification and regression training*. [https://CRAN.R-project.org/](https://CRAN.R-project.org/package=caret)
167 [package=caret](https://CRAN.R-project.org/package=caret)
- 168 Kuhn, M., & Vaughan, D. (2022). *Yardstick: Tidy characterizations of model performance*.
169 <https://CRAN.R-project.org/package=yardstick>
- 170 Kuhn, M., & Wickham, H. (2020). *Tidymodels: A collection of packages for modeling and*
171 *machine learning using tidyverse principles*. <https://www.tidymodels.org>
- 172 Lang, M., Binder, M., Richter, J., Schratz, P., Pfisterer, F., Coors, S., Au, Q., Casalicchio,
173 G., Kotthoff, L., & Bischl, B. (2019). mlr3: A modern object-oriented machine learning
174 framework in R. *Journal of Open Source Software*. <https://doi.org/10.21105/joss.01903>
- 175 Legates, D. R., & McCabe Jr., G. J. (1999). Evaluating the use of “goodness-of-fit” measures in
176 hydrologic and hydroclimatic model validation. *Water Resources Research*, 35(1), 233–241.
177 [https://doi.org/https://doi.org/10.1029/1998WR900018](https://doi.org/10.1029/1998WR900018)
- 178 Lüdtke, D., Ben-Shachar, M. S., Patil, I., Waggoner, P., & Makowski, D. (2021). performance:
179 An R package for assessment, comparison and testing of statistical models. *Journal of*
180 *Open Source Software*, 6(60), 3139. <https://doi.org/10.21105/joss.03139>
- 181 Miguez, F. (2022). *Apsimx: Inspect, read, edit and run 'APSIM' "next generation" and*
182 *'APSIM' classic*. <https://CRAN.R-project.org/package=apsimx>
- 183 Nash, J. E., & Sutcliffe, J. V. (1970). River flow forecasting through conceptual models part i
184 — a discussion of principles. *Journal of Hydrology*, 10(3), 282–290. [https://doi.org/https://doi.org/10.1016/0022-1694\(70\)90255-6](https://doi.org/10.1016/0022-1694(70)90255-6)
185
- 186 Olsen, L. R., & Zachariae, H. B. (2021). *Cvms: Cross-validation for model selection*. [https://](https://CRAN.R-project.org/package=cvms)
187 CRAN.R-project.org/package=cvms
- 188 R Core Team. (2021). *R: A language and environment for statistical computing*. R Foundation
189 for Statistical Computing. <https://www.R-project.org/>
- 190 Reshef, D. N., Reshef, Y. A., Finucane, H. K., Grossman, S. R., McVean, G., Turnbaugh, P. J.,
191 Lander, E. S., Mitzenmacher, M., & Sabeti, P. C. (2011). Detecting novel associations in
192 large data sets. *Science*, 334(6062), 1518–1524. <https://doi.org/10.1126/science.1205438>
- 193 Robinson, W. S. (1957). The statistical measurement of agreement. *American Sociological*
194 *Review*, 22(1), 17–25. <http://www.jstor.org/stable/2088760>
- 195 Robinson, W. S. (1959). The geometric interpretation of agreement. *American Sociological*
196 *Review*, 24(3), 338–345. <http://www.jstor.org/stable/2089382>
- 197 Seibold, S. A. D., Heidi AND Czerny. (2022). Correction: A computational reproducibility
198 study of PLOS ONE articles featuring longitudinal data analyses. *PLOS ONE*, 17(5), 1–1.
199 <https://doi.org/10.1371/journal.pone.0269047>
- 200 Smith, E. P., & Rose, K. A. (1995). Model goodness-of-fit analysis using regression and related
201 techniques. *Ecological Modelling*, 77(1), 49–64. [https://doi.org/https://doi.org/10.1016/](https://doi.org/10.1016/0304-3800(93)E0074-D)
202 [0304-3800\(93\)E0074-D](https://doi.org/10.1016/0304-3800(93)E0074-D)
- 203 Székely, G. J., Rizzo, M. L., & Bakirov, N. K. (2007). Measuring and testing dependence by
204 correlation of distances. *The Annals of Statistics*, 35(6), 2769–2794. [https://doi.org/10.](https://doi.org/10.1214/009053607000000505)
205 [1214/009053607000000505](https://doi.org/10.1214/009053607000000505)
- 206 Tedeschi, L. O. (2006). Assessment of the adequacy of mathematical models. *Agricultural*
207 *Systems*, 89(2), 225–247. [https://doi.org/https://doi.org/10.1016/j.agsy.2005.11.004](https://doi.org/10.1016/j.agsy.2005.11.004)
- 208 Thieme, N. (2018). R generation. *Significance*, 15(4), 14–19. [https://doi.org/10.1111/j.](https://doi.org/10.1111/j.1740-9713.2018.01169.x)
209 [1740-9713.2018.01169.x](https://doi.org/10.1111/j.1740-9713.2018.01169.x)
- 210 Wallach, D., Makowski, D., Jones, J. W., & Brun, F. (2019). Chapter 9 - model evaluation. In D.
211 Wallach, D. Makowski, J. W. Jones, & F. Brun (Eds.), *Working with dynamic crop models*

- 212 (third edition) (Third Edition, pp. 311–373). Academic Press. <https://doi.org/https://doi.org/10.1016/B978-0-12-811756-9.00009-5>
- 213
- 214 Warton, D. I., Wright, I. J., Falster, D. S., & Westoby, M. (2006). Bivariate line-fitting
215 methods for allometry. *Biological Reviews*, 81(2), 259–291. <https://doi.org/https://doi.org/10.1017/S1464793106007007>
- 216
- 217 Wickham, H. (2016). *ggplot2: Elegant graphics for data analysis*. Springer-Verlag New York.
218 ISBN: 978-3-319-24277-4
- 219 Wickham, H., Averick, M., Bryan, J., Chang, W., McGowan, L. D., François, R., Golem, G.,
220 Haye, A., Henry, L., Hester, J., Kuhn, M., Lapeere, M., Mille, E., Bach, S. M.,
221 Müller, K., Roldán, J., Sagan, M., Sievert, C., Taneer, M., Thrun, M., ... Hiroaki Yutani.
222 (2019). Welcome to the tidyverse. *Journal of Open Source Software*, 4(43), 1686. <https://doi.org/10.21105/joss.01686>
- 223
- 224 Willmott, C. J. (1981). ON THE VALIDATION OF MODELS. *Physical Geography*, 2(2),
225 184–194. <https://doi.org/10.1080/02723646.1981.10642213>
- 226 Willmott, C. J., Ackleson, S. G., Davis, R. E., Feddes, J. J., Klink, K. M., Legates, D. R.,
227 O'Donnell, J., & Rowe, C. M. (1985). Statistics for the evaluation and comparison of models.
228 *Journal of Geophysical Research: Oceans*, 90(C5), 8995–9005. <https://doi.org/https://doi.org/10.1029/JC090iC05p08995>
- 229
- 230 Willmott, C. J., Robeson, S. M., & Matsuura, K. (2012). A refined index of model performance.
231 *International Journal of Climatology*, 32(13), 2088–2094. <https://doi.org/https://doi.org/10.1002/joc.2419>
- 232
- 233 Yang, J. M., Yang, J. Y., Liu, S., & Hoogenboom, G. (2014). An evaluation of the statistical
234 methods for testing the performance of crop models with observed data. *Agricultural*
235 *Systems*, 127, 81–89. <https://doi.org/https://doi.org/10.1016/j.agsy.2014.01.008>
- 236 Zambrano-Bigiarini, M. (2020). *hydroGOF: Goodness-of-fit functions for comparison of sim-*
237 *ulated and observed hydrological time series*. <https://doi.org/10.5281/zenodo.839854>