

# Simple Box-Cox probabilistic models for hourly streamflow predictions

Cristina Prieto<sup>1,2,3</sup>, Dmitri Kavetski<sup>4,1</sup>, Fabrizio Fenicia<sup>3</sup>, James Kirchner<sup>2,5,6</sup>, David McInerney<sup>4</sup>, Mark Thyer<sup>4</sup>, and César Álvarez<sup>1</sup>

5 <sup>1</sup>IHCantabria—Instituto de Hidráulica Ambiental de la Universidad de Cantabria, Santander, Spain

<sup>2</sup>Department of Environmental Systems Science, ETH Zürich, Zürich, Switzerland

<sup>3</sup>Eawag, Swiss Federal Institute of Aquatic Science and Technology, Dübendorf, Switzerland

<sup>4</sup>School of Civil, Environmental and Mining Engineering, University of Adelaide, Adelaide, SA, Australia

<sup>5</sup>Swiss Federal Research Institute WSL, Birmensdorf, Switzerland

10 <sup>6</sup>Department of Earth and Planetary Science, University of California, Berkeley, California, USA

*Correspondence to:* Cristina Prieto (prietoc@unican.es)

This supplementary material details the calculation of the best-metric residual error model and the analysis of statistical significance of differences in the performance metrics of the competing residual error models.

We look at performance rankings of the residual error models.

15 We use statistical tests to assess whether differences in the performance metrics of competing residual error models are significant. For example, to compare the reliability (R) of three residual error models A, B, C and D.

- Residual error model A uses the ARMA model AR3 and the heteroscedastic transformation Log
- Residual error model B uses AR3 and BC02
- Residual error model C uses AR3 and BC05
- Residual error model D uses AR3 and SLS
- Suppose we are comparing the reliability of schemes A, B, C and D using five catchments, with respective metric values:

- R residual error model A for catchment 1: R\_AR3\_Log\_c1
- R residual error model A for catchment 2: R\_AR3\_Log\_c2
- R residual error model A for catchment 3: R\_AR3\_Log\_c3
- R residual error model A for catchment 4: R\_AR3\_Log\_c4
- R residual error model A for catchment 5: R\_AR3\_Log\_c5

- R residual error model B for catchment 1: R\_AR3\_BC02\_c1
- R residual error model B for catchment 2: R\_AR3\_BC02\_c2
- R residual error model B for catchment 3: R\_AR3\_BC02\_c3
- R residual error model B for catchment 4: R\_AR3\_BC02\_c4
- R residual error model B for catchment 5: R\_AR3\_BC02\_c5
- Similar for residual error model C and D, so that:

- R\_AR3\_BC05\_c1
- R\_AR3\_BC05\_c2
- R\_AR3\_BC05\_c3
- R\_AR3\_BC05\_c4
- R\_AR3\_BC05\_c5

45

- R\_AR3\_SLS\_c1
- R\_AR3\_SLS\_c2
- R\_AR3\_SLS\_c3
- R\_AR3\_SLS\_c4
- R\_AR3\_SLS\_c5

- For precision (P) and Volumetric Bias (VB) we do same as for R

50

- Steps:

1) identify the residual error model with the best (lowest) median metric values.

55

- If median A < median B < median C < median D, i.e., residual error model A is the “best median” residual error model.
- We apply the paired Wilcoxon signed-rank test [Bauer, 1972] to check for statistically significant differences at the 95% confidence level between
  - A vs B, A vs C, A vs D.
- The use of a paired test ensures that the metric values are compared case-by-case, i.e., R\_AR3\_Log\_c1 is only compared to R\_AR3\_BC02\_c1, R\_AR3\_BC05\_c1 and R\_AR3\_SLS\_c1, and so forth
- The statistical comparison of residual error models is carried out separately for the three-performance metrics (R, P, VB)
  - for a given metric, we establish whether the degradation in performance incurred by a residual error model other than the best-median residual error model is statistically significant.

60

2) establish the set of “best-metric” residual error model as the residual error models that are statistically similar to the best median scheme in the particular performance metric

65