Improving standard practices for prediction in ungauged basins: Bayesian approach

Cristina Prieto (1,2), Nataliya Le Vine (2), Eduardo García (1), and Raúl Medina (1)

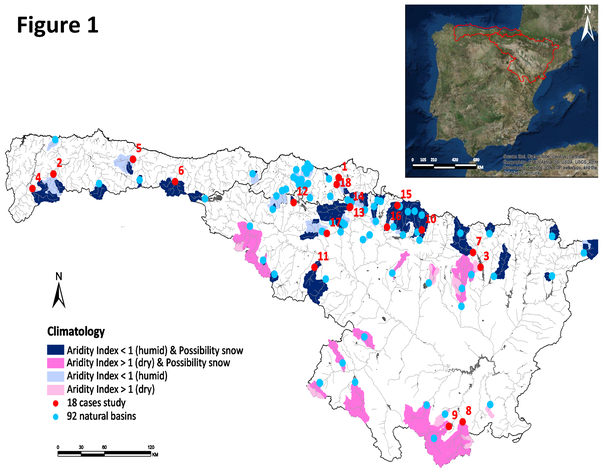
(1) IHCantabria, University of Cantabria, Santander, Spain (prietoc@unican.es), (2) Imperial College London, UK

Abstract

In hydrological modelling, prediction of flows in ungauged basins is still a defiance. A Bayesian framework based on regionalized signatures has proven to be advantageous for the task. This requires three critical assumptions about 1) selection of signatures, 2) selection of regionalization model and 3) a 'perfect' model/ input assumption. The contribution of this research is to address these hypothesis and improving the standard practices. First, a compact orthogonal set of information is selected via Principal Component Analysis. Second, functional form-assumption-free Random Forests are used to regionalize the selected information. And third, statistical metrics are used to flag-up input/model/sampling inadequacy: Bayes' Factor and 'suitability' test. The methodological developments are illustrated for flow prediction in Northern Spain.

Catchments and data

Figure 1



A dataset includes 92 natural basins with long daily flow records.

A subset of 18 basins also has long precipitation and PET time series (>8 years).

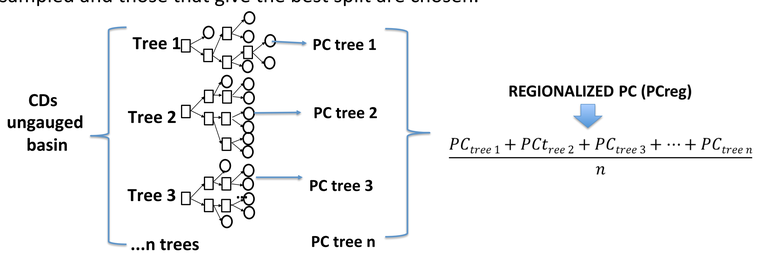
Methodology

1. Hydrological indices and catchment descriptors

103 Hydrological indices are transformed into uncorrelated principal components (PCs) by PCA. The most significant PCs are selected using the 'broken stick' method. From 50 available catchment descriptors (CDs), the least correlated 16 CDs are selected to build Random Forests regionalization models.

2. Random Forests (RFs)

RF algorithm resamples with replacement a number of regression trees to compose the forest. Each tree is grown such that at each its node a number of CDs is randomly sampled and those that give the best split are chosen.



3. Model parameter restriction - Bayes' law

$$p(\theta|PC^{reg}, M, I) = \frac{L\left(PC^{sim}_{\theta,M,I}|PC^{reg}\right) * p(\theta|M,I)}{p(PC^{reg}|M,I)}$$

4. Probabilistic Nash-Sutcliffe efficiency

Accuracy — Precision

$$NS^{prob} = \left\{1 - \frac{\sum_{t=1}^{T}\left(E[Q_t] - q_t^{obs}\right)^2}{\sum_{t=1}^{T}\left(q_t^{obs} - E[q^{obs}]\right)^2}\right\} \frac{\sum_{t=1}^{T} Var[Q_t]}{\sum_{t=1}^{T}\left(q_t^{obs} - E[q^{obs}]\right)}$$

5. Hydrological model adequacy

A) Bayes' Factor (BF)

BF quantifies how much information a model adds over regionalization. Equivalently, BF quantifies how likely a model is to represent the basin behaviour.

$$BF = \frac{P\left(PC^{data}|PC^{reg}, M, I\right)}{P\left(PC^{data}|PC^{reg}\right)} = \frac{P\left(PC^{reg}|PC^{data}, M, I\right) * P\left(PC^{data}|M,I\right)}{P\left(PC^{reg}|M,I\right) * P\left(PC^{data}|PC^{reg}\right)} = \frac{P\left(PC^{data}|M,I\right)}{P\left(PC^{data}\right)}$$

B) Model 'suitability' test

H_0={model/input/parameter sampling are able to approximate basin's PCs}

The critical elliptical regions with respect to the basin's PCs are calculated using the Mahalanobis distance. p-value for the H_0 is the smallest of the p-values for each of the critical ellipses.

Results

Four PCs explain 87.6% of variability in total (63.3%, 12.4%, 6.9%, and 4.8% of variability. Figure 2 shows residual distribution for the RFs regionalization that defines the likelihood function in the Bayes' law.

Inputs/model/sampling are tested to be adequate for only three basins (BF>1 and p-value for the suitability test >0.05) – Figure 3 shows posterior prediction bounds for one of these basins and for a 'model error' basin. NS^{prob} for the observed data and two regionalization qualities: original and 5% precision is shown on Figure 4. Figure 5 shows NS^{prob} for the 'no model error' case.

Figure 2

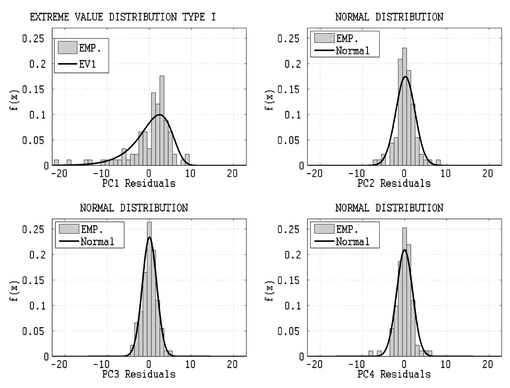


Figure 3

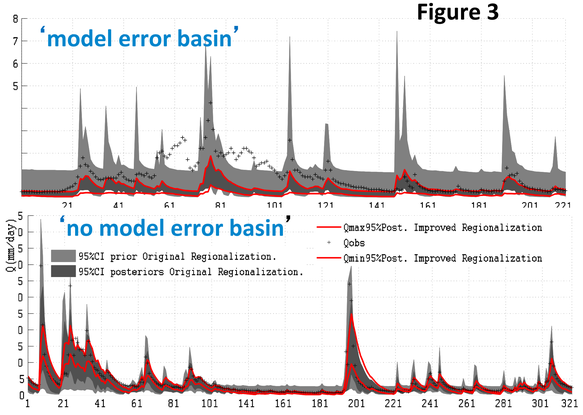


Figure 4: NS^{prob} for the observed data

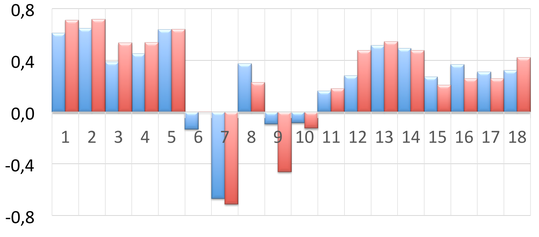
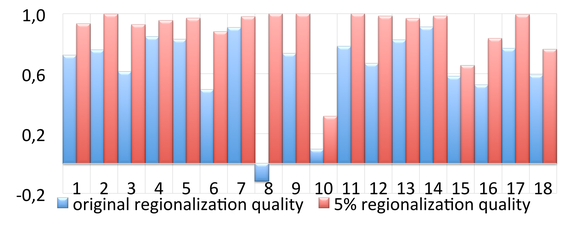


Figure 5: NS^{prob} for the synthetic data



Conclusions

- The reduced signature space provides sufficient information for considerable improvements in hydrological prediction accuracy and precision, when there is no model error.
- For a 'perfect' model, improving regionalization quality can be targeted as means for securing better predictions.
- Model error can drastically deteriorate prediction quality, so that the regionalization quality is not directly related to the prediction quality.
- The study provides some statistical metrics to quantify input/model/sampling adequacy for a given basin.