Quantification of predictive uncertainty in hydrological modelling by harnessing the wisdom of the crowd: A large-sample experiment at monthly timescale

Quantification of predictive uncertainty in hydrological modelling by harnessing the wisdom of the crowd: A large-sample experiment at monthly timescale

Georgia Papacharalampous[1,*], Hristos Tyralis[2], Demetris Koutsoyiannis[3], and Alberto Montanari[4]

[1] Department of Water Resources and Environmental Engineering, School of Civil Engineering, National Technical University of Athens, Heroon Polytechneiou 5, 157 80 Zographou, Greece; papacharalampous.georgia@gmail.com; https://orcid.org/0000-0001-5446-954X

[2] Air Force Support Command, Hellenic Air Force, Elefsina Air Base, 192 00 Elefsina, Greece; montchrister@gmail.com; https://orcid.org/0000-0002-8932-4997

[3] Department of Water Resources and Environmental Engineering, School of Civil Engineering, National Technical University of Athens, Heroon Polytechneiou 5, 157 80 Zographou, Greece; dk@itia.ntua.gr; https://orcid.org/0000-0002-6226-0241

[4] Department of Civil, Chemical, Environmental and Materials Engineering (DICAM), University of Bologna, via del Risorgimento 2, 40136 Bologna, Italy; alberto.montanari@unibo.it; https://orcid.org/0000-0001-7428-0410

* Correspondence: papacharalampous.georgia@gmail.com, tel: +30 69474 98589

This is the accepted manuscript of an article published in Advances in Water Resources. Please cite the article as: Papacharalampous GA, Tyralis H, Koutsoyiannis D, Montanari A (2020) Quantification of predictive uncertainty in hydrological modelling by harnessing the wisdom of the crowd: A large-sample experiment at monthly timescale. Advances in Water Resources 136:103470. https://doi.org/10.1016/j.advwatres.2019.103470

Abstract: Predictive hydrological uncertainty can be quantified by using ensemble methods. If properly formulated, these methods can offer improved predictive performance by combining multiple predictions. In this work, we use 50-year-long monthly time series observed in 270 catchments in the United States to explore the performances provided by an ensemble learning post-processing methodology for issuing probabilistic hydrological predictions. This methodology allows the utilization of flexible quantile regression models for exploiting information about the hydrological model's error. Its key differences with respect to basic two-stage hydrological post-

processing methodologies using the same type of regression models are that (a) instead of a single point hydrological prediction it generates a large number of "sister predictions" (yet using a single hydrological model), and that (b) it relies on the concept of combining probabilistic predictions via simple quantile averaging. A major hydrological modelling challenge is obtaining probabilistic predictions that are simultaneously reliable and associated to prediction bands that are as narrow as possible; therefore, we assess both these desired properties of the predictions by computing their coverage probabilities, average widths and average interval scores. The results confirm the usefulness of the proposed methodology and its larger robustness with respect to basic two-stage post-processing methodologies. Finally, this methodology is empirically proven to harness the "wisdom of the crowd" in terms of average interval score, i.e., the average of the individual predictions combined by this methodology scores no worse –usually better– than the average of the scores of the individual predictions.

Key words: ensemble learning; hydrological model; probabilistic prediction; quantile averaging; quantile regression; uncertainty quantification

1. Introduction

Uncertainty is a subject of ongoing discussions in hydrology (see e.g., Beven 1993, 2000, 2001; Vogel 1999; Beven and Feer 2001; Krzysztofowicz 2001a; Pappenberger and Beven 2006; Koutsoyiannis and Montanari 2007; Montanari 2007; Koutsoyiannis et al. 2009; Koutsoyiannis 2010, 2011; Kuczera et al. 2010; Ramos et al. 2010, 2013; Weijs et al. 2010; Juston et al. 2012; Nearing at al. 2016). Hydrological modelling uncertainty is traditionally recognised within the model calibration and validation phases (Montanari 2011) in the context of the widely accepted evaluation framework proposed by Klemeš (1986). Within this framework "uncertainty treatment" serves the verification of hydrological model's reliability (Montanari 2011). The large number of relevant studies and their high significance are summarised, for instance, in the review papers by Efstratiadis and Koutsoyiannis (2010), and Pechlivanidis et al. (2011).

As discussed in Koutsoyiannis (2010), an appropriate modelling approach for any uncertain hydrological system should necessarily include quantification of its uncertainty within a stochastic framework. Uncertainty is naturally quantified using the

probability theory, i.e., in terms of probability distribution function (PDF; Todini 2007; see also Todini 2004, 2008). Todini (2007; quoting Krzysztofowicz 1999) emphasizes the fact that in engineering applications the targeted uncertainty quantification should be no other than the quantification of the predictive uncertainty, i.e., the total uncertainty of the predictand. Along with this strong engineering-oriented interest of hydrologists (which might be underestimated in some cases but is of vital significance for hydrology, as for any applied science; Shmueli 2010), understanding of predictive performance and uncertainty in hydrological modelling is undoubtedly a major science-oriented target (see e.g., Clark et al. 2008; Renard et al. 2010, 2011; Montanari 2011; Pechlivanidis et al. 2011; Beven 2012; Montanari and Koutsoyiannis 2012; Clark et al. 2015; Farmer and Vogel 2016; Széles et al. 2018; Khatami et al. 2019).

The preference for process-based (including conceptual) hydrological models (over the data-driven ones; Toth et al. 1999), along with both the practical relevance of predictive uncertainty quantification in hydrology and the attentiveness of hydrologists towards increasing understanding in (probabilistic) hydrological modelling, has led to the development of a wide range of methodologies for the integration of process-based and statistical models. This range includes (but is not limited to) various types of methodologies that statistically post-process the output of process-based models (hereafter referred to as "post-processing" methodologies). Considering information from deterministic models within uncertainty assessment frameworks (instead of exclusively using statistical methods) is a state-of-the-art methodological approach that is also adopted in contiguous fields (see e.g., Tyralis and Koutsoyiannis 2017). This approach holds a prominent position in the field of probabilistic hydrological modelling, in contrast to purely statistical probabilistic methodologies, which are rarely preferred; therefore, the below-provided outline exclusively focuses on it.

Perhaps the most frequently exploited methodology for predictive uncertainty quantification in hydrological modelling is the Generalized Likelihood Uncertainty Estimation (GLUE; Beven and Binley 2014). This approach has been proposed by Beven and Binley (1992), and is based on the concept of equifinality (see, e.g., Beven 2006; Khatami 2019). It has been discussed, for example, in Montanari (2005), Mantovan and Todini (2006), Stedinger et al. (2008), Vrugt et al. (2009b), and Sadegh and Vrugt (2013); see also the related comments in Todini (2007).

Another predictive uncertainty quantification methodology that has received

attention both by researchers and practitioners is the Bayesian Forecasting System (BFS). The BFS has been introduced by Krzysztofowicz (1999, 2001a, 2002), Krzysztofowicz and Kelly (2000), and Krzysztofowicz and Herr (2001) for producing probabilistic river stage forecasts. It consists of three discrete components, namely the Precipitation Uncertainty Processor (PUB), the Hydrologic Uncertainty Processor (HUP) and the INTegrator (INT). Information about these components can be found in Kelly and Krzysztofowicz (2000), Krzysztofowicz and Kelly (2000), and Krzysztofowicz (2001b) respectively. This Bayesian methodology is conceived for real-time forecasting and relies on the assumption that uncertainty is mainly introduced by rainfall forecast errors.

There are also Bayesian post-processing methodologies that explicitly consider the contribution of input and output data uncertainty (which also affects the quantification of parameter uncertainty; see Coxon et al. (2015), Di Baldassarre et al. (2012), Di Baldassarre and Montanari (2009), Kauffeldt et al. (2013), McMillan et al. (2010), McMillan et al. (2012), Montanari and Di Baldassarre (2013), and Tomkins (2014) for information on rainfall-runoff data errors). Perhaps the most characteristic example of such a methodology is the Bayesian Total Error Analysis (BATEA) framework by Kavetski et al. (2002; see also Kavetski et al. 2006a, Kuczera et al. 2006), implemented, for instance, in Thyer et al. (2009) and Renard et al. (2010, 2011). This Bayesian framework facilitates the joint modelling of parameter uncertainty, data uncertainties, and model error, i.e., of all sources of uncertainty that are often assumed to collectively compose the predictive uncertainty. Other Bayesian post-processing methodologies introduced for parameter and predictive uncertainty quantification are described by Kuczera (1983), Schoups and Vrugt (2010), Evin et al. (2013; see also Evin et al. 2014), Hernández-López and Francés (2017) and Romero-Cuellar et al. (2019); see also the literature review in Hernández-López and Francés (2017).

Non-Bayesian post-processing methodologies that in their majority focus on the modelling of a single error term conditional on hydrological point predictions and historical information are also available in the hydrological modelling literature (see e.g., Bock et al. 2018; Bourgin et al. 2015; Farmer and Vogel 2016; Montanari and Brath 2004; Montanari and Grossi 2008; Dogulu et al. 2015; López López et al. 2014; Solomatine and Shrestha 2009; Wani et al. 2017). Adopting the terminology by Evin et al. (2014), such methodologies are hereafter referred to as “two-stage” post-processing

methodologies, as their hydrological and error models are estimated in two subsequent stages. It is relevant to note at this point that Bayesian and two-stage post-processing methodologies are rather not directly comparable, since they are characterized by different statistical-modelling-culture traits and distinguishing features, which in their turn lead to different advantages and disadvantages (see Appendix A). For extensive discussions on the statistical modelling cultures, the reader is referred to Breiman (2001) and Shmueli (2010).

In the context described so far, Montanari and Koutsoyiannis (2012) introduced a flexible two-stage post-processing methodology (hereafter referred to as "MK blueprint methodology") that facilitates both probabilistic modelling and understanding from a stochastic perspective of rainfall-runoff (and other stochastic) relationships. In its basic configuration, this methodology utilizes a single hydrological model to generate a large number of point predictions (hereafter referred to as "sister predictions"; adopting a similar terminology to the one by Nowotarski et al. 2016, Wang et al. 2016, and Liu et al. 2017). As implied by its post-processing nature, it also utilizes a second –necessarily statistical– model for modelling the error of the hydrological model (hereafter referred to as "error model").

Different variants of the MK blueprint methodology can be found in Sikorska et al. (2015), Quilty et al. (2019) and Papacharalampous et al. (2019b; companion to the present paper). The original blueprint and the variant by Sikorska et al. (2015) are formulated to explicitly consider input data uncertainty, while in both related papers a large number of hydrological model parameters are obtained by using the DREAM algorithm by Vrugt et al. (2009a; see also Vrugt 2016). This algorithm (see, e.g., Schoups and Vrugt 2010; Laloy and Vrugt 2012; Vrugt et al. 2013; Sadegh and Vrugt 2014) is a popular Markov chain Monte Carlo (MCMC) algorithm for sampling from the posterior parameter distribution of hydrological models (see also the related implementations in Sadegh et al. 2015; Hernández-López and Francés 2017; Vrugt et al. 2008; Volpi et al. 2017). Other (non-Bayesian) methodologies could also be used for obtaining a large number of hydrological model parameters (Montanari and Koutsoyiannis 2012), while in absence of relevant information the MK blueprint methodology can also be applied without explicitly considering input data uncertainty (see e.g., the implementations in Quilty et al. 2019 and the formulations of the variants in Papacharalampous et al. 2019b). Quilty et al. (2019) perform probabilistic water demand forecasting using

exogenous variables; therefore, their variants constitute integrations within the MK blueprint framework of concepts particularly useful and/or popular for this task, such as bootstrapping, variable selection and wavelet decomposition.

In spite of their (larger or smaller) differences in terms of conceptualization, underlying modelling cultures and inherent modelling assumptions, all the above-outlined state-of-the-art techniques aim at filling a common knowledge gap that currently exists in the probabilistic hydrological modelling and forecasting literatures, specifically at answering the following research question: How to reduce modelling uncertainty as much as possible? Risk reduction in (probabilistic) hydrological modelling is the 20th of the 23 major "unsolved" hydrological problems, as posed by Blöschl et al. (2019, Section 3) through a community-based process. The present study aspires to contribute to the large efforts made towards solving this problem.

We extensively test the hydrological modelling capabilities provided by the variants of the MK blueprint methodology introduced in Papacharalampous et al. (2019b) (hereafter collectively referred to as "working methodology"), when these variants are applied using the quantile regression model by Koenker and Bassett (1978; see also Koenker 2005) as error model. The quantile regression model is a balanced choice between interpretable and more flexible algorithms from the statistical learning literature. It has already been applied for post-processing hydrological predictions within hydrological modelling case studies (see e.g., Dogulu et al. 2015, López López et al. 2014, Solomatine and Shrestha 2009, Wani et al. 2017), while its use is more common in the field of hydrological forecasting (see e.g., Tyralis et al. 2019a and the references therein); see also the references in Dogulu et al. (2015), and Abbas and Xuan (2019) for applications of this model in other geoscience concepts.

For benchmarking purposes, we also apply the working methodology using the linear regression model (see e.g., James et al. 2013; Hastie et al. 2009) as error model, and two naïve probabilistic data-driven schemes. For the merits of using benchmarks in hydrological modelling, the reader is referred to Pappenberger et al. (2015); see also benchmarking examples in Montanari and Brath (2004), Papacharalampous and Tyralis (2018), Papacharalampous et al. (2018a,b,c, 2019a,b,d), Quilty et al. (2019), Evin et al. (2014), Sikorska et al. (2015), Tyralis and Papacharalampous (2017, 2018), Tyralis et al. (2018, 2019a,c), and Xu et al. (2018).

The working methodology is implemented within a large-sample real-world experiment. In the latter, we probabilistically solve monthly rainfall-runoff modelling problems for 270 catchments in the United States (US). Large-sample hydrological studies are increasingly carried out in the literature (see e.g., Bock et al. 2018; Bourgin et al. 2015; Coxon et al. 2015; Farmer and Vogel 2016; Langousis et al. 2016; Mouelhi et al. 2006a,b; Papacharalampous et al. 2018a,b, 2019a,d; Papalexiou and Koutsoyiannis 2013; Perrin et al. 2001; Ren et al. 2016; Sawicz et al. 2011; Tyralis and Koutsoyiannis 2017; Tyralis and Papacharalampous 2017, 2018; Tyralis et al. 2018, 2019a,c; Weijs et al. 2013; Xu et al. 2018, 2019), while this is the first study performing a large-scale assessment of the MK blueprint methodology.

The aims of the study (that can be addressed only within a large-sample hydrological study) are to:

1) Validate the working methodology.
2) Compare its variants both in terms of predictive performance and computational requirements.
3) Quantify the improvement in performance when using the quantile regression model instead of the linear regression model as error model. In contrast to the latter model, the former model is known to be appropriate for modelling heteroscedasticity (Koenker and Hallock 2001; Koenker 2005).
4) Demonstrate in real-world applications the larger robustness in performance of the working methodology compared to two-stage post-processing methodologies producing a single point hydrological prediction (hereafter referred to as "basic" two-stage post-processing methodologies).
5) Provide an empirical proof of the ability of the working methodology to harness the wisdom of the crowd. This ability stems from the concept of combining probabilistic predictions via simple quantile averaging, on which this methodology relies, while in Lichtendahl et al. (2013, Section 5) it is defined as follows: The average of predictions scores no worse –usually better– than the average of the scores of the combined predictions. According to the same study, this ability has to be empirically proven for the problem and scores of interest, since the proofs in Lichtendahl et al. (2013) are made for stylized versions.

2. Data and methods

In this section, we present the experimental methodology of the study by emphasizing implementation details, as it is suggested by the guidelines by Abrahart et al. (2008). Statistical software information is summarized in Appendix B. The working methodology is outlined in Appendix C, while the reader is referred to Papacharalampous et al. (2019b) for its detailed and formal presentation.

2.1 Rainfall-runoff dataset

We use the US Model Parameter Estimation Experiment (MOPEX) dataset, which is documented in Schaake et al. (2006; see also Schaake et al. 2000, Duan et al. 2006, Wagener et al. 2006). This dataset comprises hydrometeorological and land-surface-characteristic data originating from US catchments of intermediate size, and has been extensively used in hydrological studies (see e.g., Kavetski et al. 2006b; Sawicz et al. 2011; Huang et al. 2013; Evin et al. 2014; Weijs et al. 2013; Ye et al. 2014; Ren et al. 2016; Hernández-López and Francés 2017). All included catchments are unregulated; therefore, the modelling assumption of stationarity is reasonable on these real-world data (see e.g., Koutsoyiannis 2011; Montanari and Koutsoyiannis 2014; Koutsoyiannis and Montanari 2015).

From the original dataset we retrieve daily information about mean areal precipitation, climatic potential evaporation and streamflow discharge for 431 US catchments. The retrieved data span from January 1st, 1948 to December 31st, 2003, thus covering a 56-year period, yet containing a large amount of missing and negative (unrealistic) values. We process the retrieved data aiming to simultaneously achieve two objectives, i.e., (a) extracting time series blocks covering a long common period of complete historical information (with no missing or unreliable values), and (b) retaining historical information for a large number of catchments. A satisfactory compromise between these two objectives is reached when using as sampling period each of the periods 1950–1999 and 1949–1998. Both these samplings result in 50 (calendar) years of complete daily time series data for 270 catchments. We adopt the former option, as it offers (slightly) more recent data compared to the alternative one. The retained time series data are aggregated to produce total monthly precipitation, potential evaporation and streamflow discharge time series, each comprising 600 values. The resulted total monthly time series constitute the herein examined dataset. The locations of the

examined MOPEX catchments are depicted in Figure 1. A wide range of climate regimes is well-represented by this sample set of catchments (see Kottek et al. 2006).

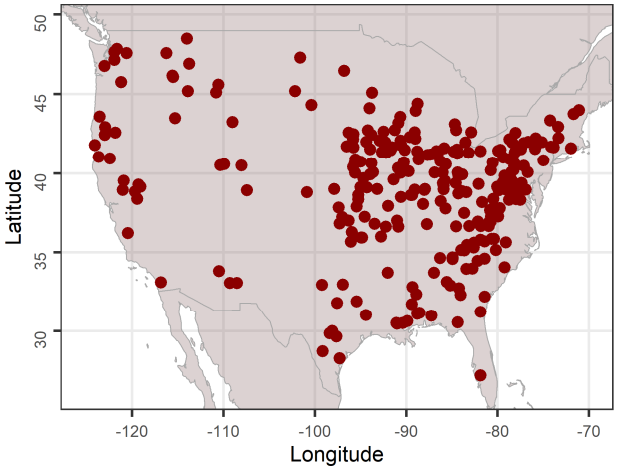

Figure 1. Locations of the 270 MOPEX catchments examined within the large-sample experiment of the study. The data are sourced from Schaake et al. (2006).

2.2 Prediction interval obtainment

2.2.1 Overview of modelling methodology

The monthly data of Section 2.1 are handled as described in Section 2.2.2. We use these data to assess two basic and six ensemble schemes in obtaining interval predictions. Two statistical learning regression models (see Section 2.2.3) and one hydrological model (see Section 2.2.4) are utilized for this assessment. We define the prediction problem to be solved as the problem of predicting the quantiles with probability $p \in \{0.005, 0.0125, 0.025, 0.05, 0.10, 0.90, 0.95, 0.975, 0.9875, 0.995\}$ of monthly streamflow discharge in the period T_3 (hereafter referred to as "quantiles of interest") given monthly precipitation and monthly potential evaporation observations for the period $\{T_0, T_1, T_2, T_3\}$ and monthly streamflow discharge observations for the period $\{T_0, T_1, T_2\}$. These periods are defined in Section 2.2.2.

The basic schemes are "linear regression" and "quantile regression". Both of them are implemented by training the regression model directly on monthly data for the period $\{T_0, T_1, T_2\}$ and, subsequently, by using the trained regression model to predict the quantiles of interest (for the period T_3). The predictor variables in regression are monthly precipitation at time t and monthly potential evaporation at time t, while the response variable is monthly streamflow discharge at time t. We note that these benchmark implementations of the regression models can only be viewed as naïve data-

driven approaches to probabilistic hydrological modelling (because of the small number of predictor variables utilized). For more sophisticated implementations (which are outside of the scope of the study), more predictor variables could be used.

On the other hand, the ensemble schemes can be perceived as different configurations of the working methodology (allowing us to address the aims of the study). Ensemble schemes 1–3 (4–6) are based on variants 1–3 respectively of this methodology. Moreover, ensemble schemes 1–3 utilize a different statistical learning regression model as error model with respect to ensemble schemes 4–6. Specifically, ensemble schemes 1–3 utilize the linear regression model, while ensemble schemes 4–6 utilize the quantile regression model. The same ensemble schemes are also implemented in Papacharalampous et al. (2019b); however, their implementation therein is made by using toy hydrological models.

We describe here below the application of the ensemble schemes for a single catchment; the extension to all catchments is straightforward. The following steps are made once for all ensemble schemes:

1) We use monthly precipitation, potential evaporation and streamflow discharge observations for the period T_1 to obtain 600 sets of the hydrological model's parameters, as detailed in Section 2.2.4. This number of parameter sets offers a good compromise between computational requirements and predictive performance. We use these parameters to define 600 sister model realizations.

2) We obtain 600 sister predictions for the period $\{T_2, T_3\}$. Each sister prediction is obtained by implementing a different sister model realization given the monthly precipitation and potential evaporation observations for the same period. Each sister prediction contains 444 values.

3) We compute the sister model realizations' errors in the period T_2 by using the parts of the sister predictions extending in the same period alongside with their corresponding target values. The total number of the computed error values is 600 × 144 = 86 400.

The following steps are made independently by each ensemble scheme:

4) We train the error model in the period T_2. Specifically, we regress the sister model realizations' error at time t (response variable) on the sister prediction at time t (predictor variable). Ensemble schemes 1 and 4 train the error model 600 times,

each time using a different sister prediction and its corresponding sister model realization's errors (use of 600 training datasets of size 144). Ensemble schemes 2 and 5 train the error model once by using all sister predictions and their corresponding sister model realizations' errors (use of one training dataset of size 86 400). Ensemble schemes 3 and 6 train the error model once by using a randomly selected sister prediction and its corresponding sister model realization's errors (use of one training dataset of size 144). The result of this step is 600 trained versions of the error model (each corresponding to a specific sister prediction) for each of the ensemble schemes 1 and 4, and one trained version of the error model for each of the ensemble schemes 2, 3, 5 and 6.

5) We apply the trained versions of the error models, obtained in the preceding step, to predict the quantiles with probability $p \in \{0.005, 0.0125, 0.025, 0.05, 0.10, 0.90, 0.95, 0.975, 0.9875, 0.995\}$ of each sister model realization's errors in the period T_3 given their corresponding sister prediction. For each ensemble scheme, the result of this step is 600 probabilistic predictions, each consisting of 10 quantile predictions.
6) We obtain 600 auxiliary probabilistic predictions of the process of interest, each consisting of 10 quantile predictions, by subtracting each of the 600 × 10 = 6 000 quantile predictions from its corresponding sister prediction.
7) The finally delivered predictive quantile with probability $p \in \{0.005, 0.0125, 0.025, 0.05, 0.10, 0.90, 0.95, 0.975, 0.9875, 0.995\}$ at time $t \in T_3$ is the average over all auxiliary predictive quantiles with the same probability p at time t, i.e., the average of 600 in number auxiliary predictive quantiles. The finally delivered predictive quantiles of the process of interest form the 99%, 97.5%, 95%, 90% and 80% central prediction intervals.

The total number of sister predictions produced herein is 270 × 600 = 162 000, each containing 444 values, while the total number of auxiliary quantile predictions is 270 × 600 × 10 × 6 = 9 720 000, each containing 300 values, and the finally delivered quantile predictions are 270 × 10 × 8 = 21 600, each containing 300 values. For addressing aim 2 of the study, we measure the computational time consumed by each ensemble scheme.

2.2.2 Data handling and related remarks

Following the notations provided in Appendix C, we define the periods $T_1 = \{13, \ldots, 156\}$, $T_2 = \{157, \ldots, 300\}$ and $T_3 = \{301, \ldots, 600\}$ (corresponding to years 1951–1962,

1963–1974 and 1975–1999 respectively). We include a large amount of the available information in the period T_3 to facilitate proper testing. We also define period T_0 = {1, …, 12} (corresponding to year 1950). This period is used for warming-up the hydrological model (see Section 2.2.4). One-year warming-up periods are often assumed adequate for achieving an optimal state initialisation, while also allowing the full exploitation of the available historical information (see e.g., Edijatno et al. 1999; Perrin et al. 2003; Kim et al. 2018; see also the implementations in Xu 2001; Perrin et al. 2001; Mouelhi et al. 2006b; Vrugt et al. 2008).

We note that the data are used without any transformation applied to it. We attempted to apply the linear regression and quantile regression schemes to river discharge data that were pre-processed by using the square-root transformation. Nevertheless, this pre-processing (not presented here for reasons of brevity) had a negative effect on the quality of the naïve probabilistic predictions, mainly to those delivered by the linear regression scheme; therefore, it was abandoned. Moreover, a logarithmic transformation was not feasible, due to some zero monthly values of river discharge. We also attempted to apply the Yeo-Johnson and ordered quantile normalization transformations on the response, when solving the error modelling problems outlined in Section 2.2.1 (steps 4–5 of the application of the ensemble schemes). These transformations were also abandoned due to infinite predicted values. The square-root and logarithmic transformations on the response variable, i.e., the error of the hydrological model at time t, are not feasible due to the existence of negative error values.

2.2.3 Regression models and related procedures

We implement the linear regression and quantile regression models. Koenker and Hallock (2001) comprehensively discuss the difference in rationale behind these two models, as summarized subsequently. The training outcome in linear regression (i.e., least-squares regression with i.i.d. Gaussian errors with zero mean and constant variance; James et al. 2013) is a conditional mean function. The latter is a function describing how the mean of the response variable changes with the changes of the predictor variables. This function is obtained by minimizing a sum of squared residuals. On the contrary, the training outcome in quantile regression is a set of conditional quantile functions, obtained by minimizing the average quantile score. While in linear

regression the PDF of the response variable is assumed to have the exact same variance and distributional shape independently of the values of the predictors, quantile regression does not make any particular assumption about this PDF; therefore, allowing a more representative description of the relationship between predictors and predictand. We use these two models to solve the regression problems described in Section 2.2.1. We train the quantile regression model by implementing the training algorithm by Koenker and d'Orey (1987, 1994).

2.2.4 Hydrological model and related procedures

We implement the monthly GR2M model by Mouelhi et al. (2006b), a parsimonious lumped conceptual model comprising only two parameters, that has been widely applied in the literature (see e.g., Paturel et al. 1995; Niel et al. 2003; Huard and Mailhot 2008; Louvet et al. 2016). This model was developed by adopting a stepwise procedure aiming to identify the most useful components of a five-parameter model. The latter was inspired from the structures of the monthly model by Makhlouf and Michel (1994), and the daily GR4J model by Perrin et al. (2003; see also Edijatno et al. 1999, Perrin et al. 2001). The first parameter (θ_1) is the maximum capacity of the soil moisture reservoir expressed in mm, while the second one (θ_2) represents water exchange between the studied and adjacent catchments. Values of the second parameter larger (smaller) than 1 indicate water supply from (to) adjacent catchment(s).

We simulate the posterior distribution of the parameters of the GR2M model conditional on the observations of the period T_1 within a Bayesian MCMC framework. We use flat priors for both the parameters θ_1 and θ_2. The likelihood error function is defined by Equation (1), where y_t is the monthly streamflow discharge observations at time t, $u_t(\theta_1, \theta_2)$ is the prediction of the GR2M model at time t and $|T_1|$ is the number of target data points included in the period T_1. We run 3 parallel Markov chains with different initial values, each comprising 2 000 iterations. The iterative simulation is performed by using the DRAM algorithm by Haario et al. (2006).

$$L(\theta_1, \theta_2) \propto \left(\sum_t \left(y_t - u_t(\theta_1, \theta_2)\right)^2\right)^{-|T_1|/2} \quad (1)$$

We assess the approximate convergence of these chains by implementing the algorithm of Brooks and Gelman (1998), i.e., a multivariate version of the algorithm of Gelman and Rubin (1992). Amongst the outputs of this algorithm is a point estimate that is assumed to be informative about the approximate convergence, while it is based on a

comparison of within-chain and between-chain variances. Point estimates substantially larger than 1 indicate lack of convergence. The simulation process is repeated until a point estimate smaller than 1.10 is delivered. Once the simulation is over, we retain the last 200 values of each chain, i.e., 600 values in total for each catchment. An example of simulated and retained parameters is presented in Figure 2.

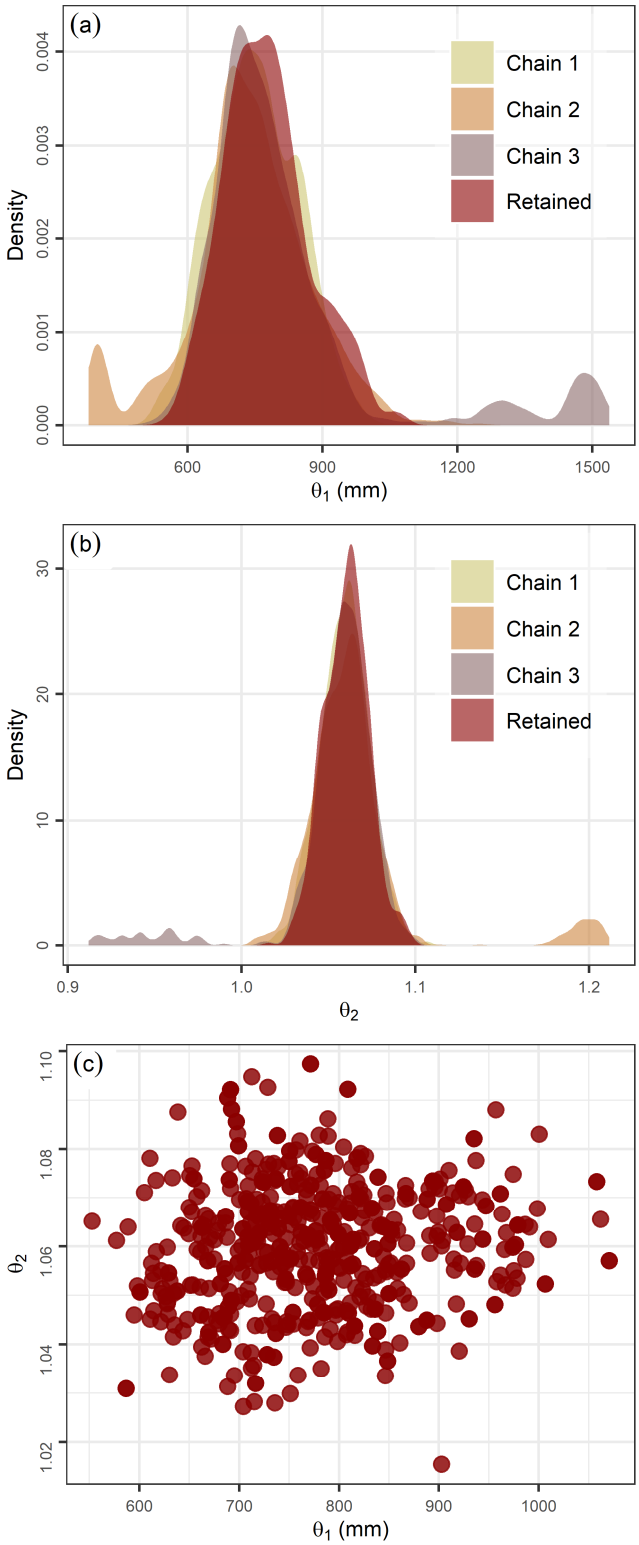

Figure 2. Simulated chains in (a–b), and retained parameter values in (a–c) obtained using precipitation, potential evaporation and streamflow discharge information for the period T_1 (years 1951–1962) for a randomly selected catchment.

2.3 Prediction interval assessment

We assess the quality of the interval predictions by computing their coverage probabilities, average widths and average interval scores. These metrics are used

according to Table 1 to assess two desired properties in probabilistic modelling, i.e., the reliability and sharpness of interval predictions. The former property is defined as the statistical correspondence between the probabilistic forecasts and the observations, while the latter is the concentration of the predictive PDFs in absolute terms (Gneiting and Katzfuss 2014; see also Gneiting et al. 2007; Gneiting and Raftery 2007). For illustrative purposes, we also present examples of prediction intervals. We do not present QQ-plots for the following two reasons: i) we deliver predictive quantiles with probabilities that are either equal or smaller than 0.10, or equal or larger than 0.90 (since we are interested in specific prediction intervals; see Section 2.2.1), while QQ-plots are ideal when PDF predictions (or at least sets of predictive quantiles with probabilities running on a grid from 0 to 1) are delivered, and ii) we are interested in objectively assessing on a massive scale the predictive performance of several prediction schemes (separately for each of them) in 270 catchments, while QQ-plots are particularly useful for assessments made on a smaller scale.

Table 1. Metrics used for assessing the prediction interval $(1 - \alpha)$, $0 < \alpha < 1$.

Metric	Definition	Possible values	Preferred values	Criterion/criteria
Coverage probability (CP_α)	Equation (2)	[0, 1]	Smaller $\lvert CP_\alpha - (1 - \alpha)\rvert$	Reliability
Average width (AW_α)	Equation (3)	$[0, +\infty)$	Smaller AW_α	Sharpness
Average interval score (AIS_α)	Equation (4)	$[0, +\infty)$	Smaller AIS_α	Reliability, sharpness

For a specific central prediction interval of level $(1 - \alpha)$, $0 < \alpha < 1$, extending in the period T_3, the coverage probabilities, average widths and average interval scores are defined with Equations (2–4) respectively, where $v_{p,t}$ is the predictive quantile with probability $p \in \{\alpha/2, 1 - \alpha/2\}$ of monthly streamflow discharge at time t, $I(\cdot)$ is the indicator function and $|T_3|$ is the number of the target data points included in the period T_3.

$$CP_\alpha := \sum_t I(y_t \in [v_{(\alpha/2),t}, v_{(1-\alpha/2),t}])/|T_3| \quad (2)$$

$$AW_\alpha := \sum_t (v_{(1-\alpha/2),t} - v_{(\alpha/2),t})/|T_3| \quad (3)$$

$$AIS_\alpha := \sum_t ((v_{(1-\alpha/2),t} - v_{(\alpha/2),t}) + (2/\alpha)(v_{(\alpha/2),t} - y_t) I(y_t < v_{(\alpha/2),t}) + (2/\alpha)(y_t - v_{(1-\alpha/2),t}) I(y_t > v_{(1-\alpha/2),t}))/|T_3| \quad (4)$$

Some remarks should be made on the (average) interval score. This score is appropriate for assessing probabilistic predictions in the form of prediction intervals (Gneiting and Raftery 2007, Section 6.2). It has three components (see Equation 4 above). The first component is the width of the prediction interval. As smaller values of the (average) interval score indicate better predictions than larger values (for a specific prediction problem), this component penalizes more the wider prediction intervals than

the narrower ones (thereby rewarding narrow prediction intervals). The two remaining components quantify the distance between each of the two predictive quantiles forming the prediction interval and the observed value, in case that the latter falls outside of the prediction interval, and penalize larger distances more than smaller distances. In general, the (average) interval score should become smaller as we move from the outer to the inner prediction intervals. The reader is referred to Gneiting and Raftery (2007, Section 6.2) for detailed information on how to interpret this score.

Since the magnitude of the average interval score largely depends on the examined dataset, we mostly base our conclusions on relative improvements in terms of average interval score. The relative improvement in terms of average interval score, obtained when using a prediction interval P_1 (provided by a predictor of interest) with respect to another prediction interval P_2 of the same level (provided by a benchmark predictor), and denoted with RI_{P_1,P_2}, is computed according to Equation (5). In this equation, AIS_{P_1} and AIS_{P_2} denote the average interval scores of prediction interval P_1 and prediction interval P_2 respectively when they are computed over the whole time series; see Equation (4).

$$\mathrm{RI}_{P_1,P_2} := (\mathrm{AIS}_{P_2} - \mathrm{AIS}_{P_1})/\mathrm{AIS}_{P_2} \tag{5}$$

Specifically, for addressing aims 1–3 of the study we compute the relative improvements provided all prediction schemes with respect to the linear regression and quantile regression schemes, and the relative improvements provided by ensemble schemes 4–6 with respect to ensemble schemes 1–3. For addressing aim 4 of the study, we use all auxiliary quantile predictions (9 720 000 in number) and the finally delivered quantile predictions (21 600 in number) to compute the relative improvements in terms of average interval score, when using the output of each ensemble scheme instead of each of the auxiliary interval predictions combined to obtain this output, according to Equation (6). In this equation, $\mathrm{AIS}_{\mathrm{OUT}}$ denotes the average interval score of the output interval prediction (obtained by using the method), $\mathrm{AIS}_{\mathrm{IN}_i}$ the average interval score of one from the auxiliary interval predictions $\{\mathrm{IN}i,\ i = 1, \ldots, 600\}$ that are averaged by the method to obtain the output interval prediction (with average interval score equal to $\mathrm{AIS}_{\mathrm{OUT}}$), and $\mathrm{RI}_{\mathrm{OUT},\mathrm{IN}_i}$ the relative improvement of interest.

$$\mathrm{RI}_{\mathrm{OUT},\mathrm{IN}_i} := (\mathrm{AIS}_{\mathrm{IN}_i} - \mathrm{AIS}_{\mathrm{OUT}})/\mathrm{AIS}_{\mathrm{IN}_i} \tag{6}$$

Finally, for addressing aim 5 of the study we use the same quantile predictions used

for addressing aim 4 to compute the relative differences between the average interval score computed for the outputs of the ensemble schemes, i.e., the average of 600 probabilistic predictions (denoted with AIS_{OUT}; see above), and the average of the average interval scores computed for each of the combined auxiliary interval predictions $\{AIS_{IN_i}, i = 1, \ldots, 600\}$ (denoted with $AAIS_{IN}$; see also Equation (7) for its definition), the latter used as reference for the former. The computation of these relative differences is made using an equation analogous to Equations (5) and (6) above, specifically Equation (8), where $RD_{OUT,AAIS_{IN}}$ denotes the relative difference of interest.

$$AAIS_{IN} := \sum_{i=1}^{600}(AIS_{IN_i})/600 \tag{7}$$

$$RD_{OUT,AAIS_{IN}} := (AAIS_{IN} - AIS_{OUT})/AAIS_{IN} \tag{8}$$

3. Results and discussions

3.1 Addressing aims 1–3 of the study

This section is devoted to addressing aims 1–3 of the study. The presentation is mostly made in an aggregated form across all the examined catchments, while emphasis is placed on the average interval scores computed for the obtained prediction intervals and on the relative improvements provided by the ensemble schemes with respect to the basic schemes in terms of the same metric. This choice is implied by the fact that an objective co-assessment regarding reliability and sharpness provided, for instance, by the interval score is of the most practical relevance in technical applications; for a justification see Papacharalampous et al. (2019b); see also Gneiting and Katzfuss (2014). In spite of this placed emphasis and keeping pace with studies, such as those of Renard et al. (2010, 2011), Evin et al. (2013, 2014), Papacharalampous et al. (2019d) and Tyralis et al. (2019a), we start the presentation by separately summarizing the information that is purely related to the assessment of reliability from the information that is purely related to the assessment of sharpness. In this way, we facilitate an adequate degree of interpretability and understanding of what follows.

In Figure 3, we present several examples of prediction intervals, all delivered by ensemble scheme 5, in comparison to the targeted data points. As extracted from Figure 3, this scheme offers a (rather) high degree of reliability, i.e., it delivers prediction intervals that mostly contain the desired percentage of data points. The same applies to the remaining prediction schemes. Herein the related information is objectively

summarized with Figure 4 and Table 2. In Figure 4, we comparatively present the boxplots of the coverage probabilities computed for all delivered and assessed solutions to the 270 examined rainfall-runoff problems. These coverage probabilities are rather good (than bad). The latter characterization holds, especially if we consider that the examined monthly time series are of only 600 values. In particular, the coverage probabilities for the 95% prediction intervals are comparable to those computed for the probabilistic predictions of Bock et al. (2018).

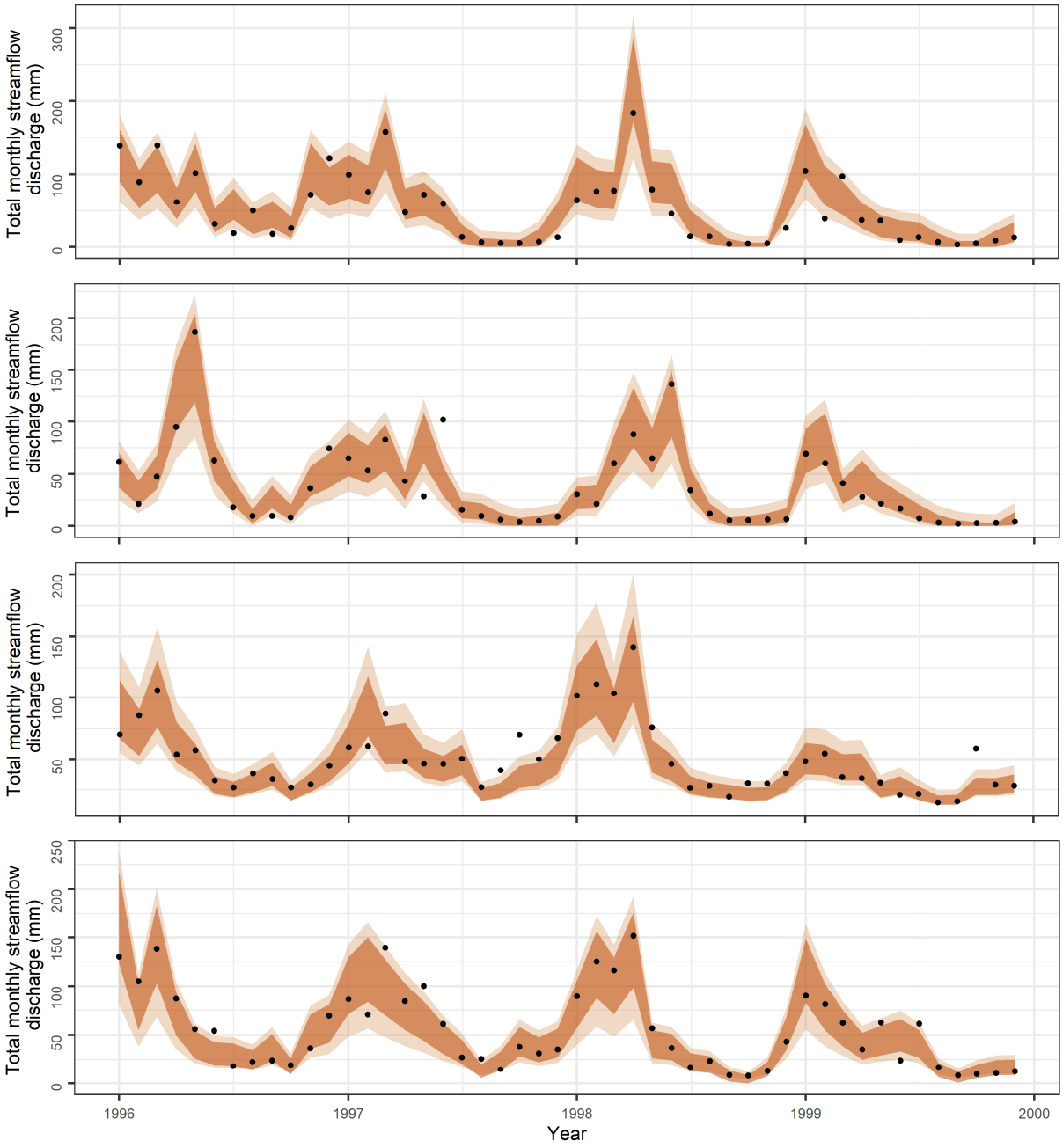

Figure 3. Prediction intervals provided by ensemble scheme 5 for four arbitrary catchments and a common 4-year sub-period of the period T_3 (years 1996–1999). Black dots denote the targeted points, while light orange and dark orange ribbons denote the 95% and 80% prediction intervals respectively.

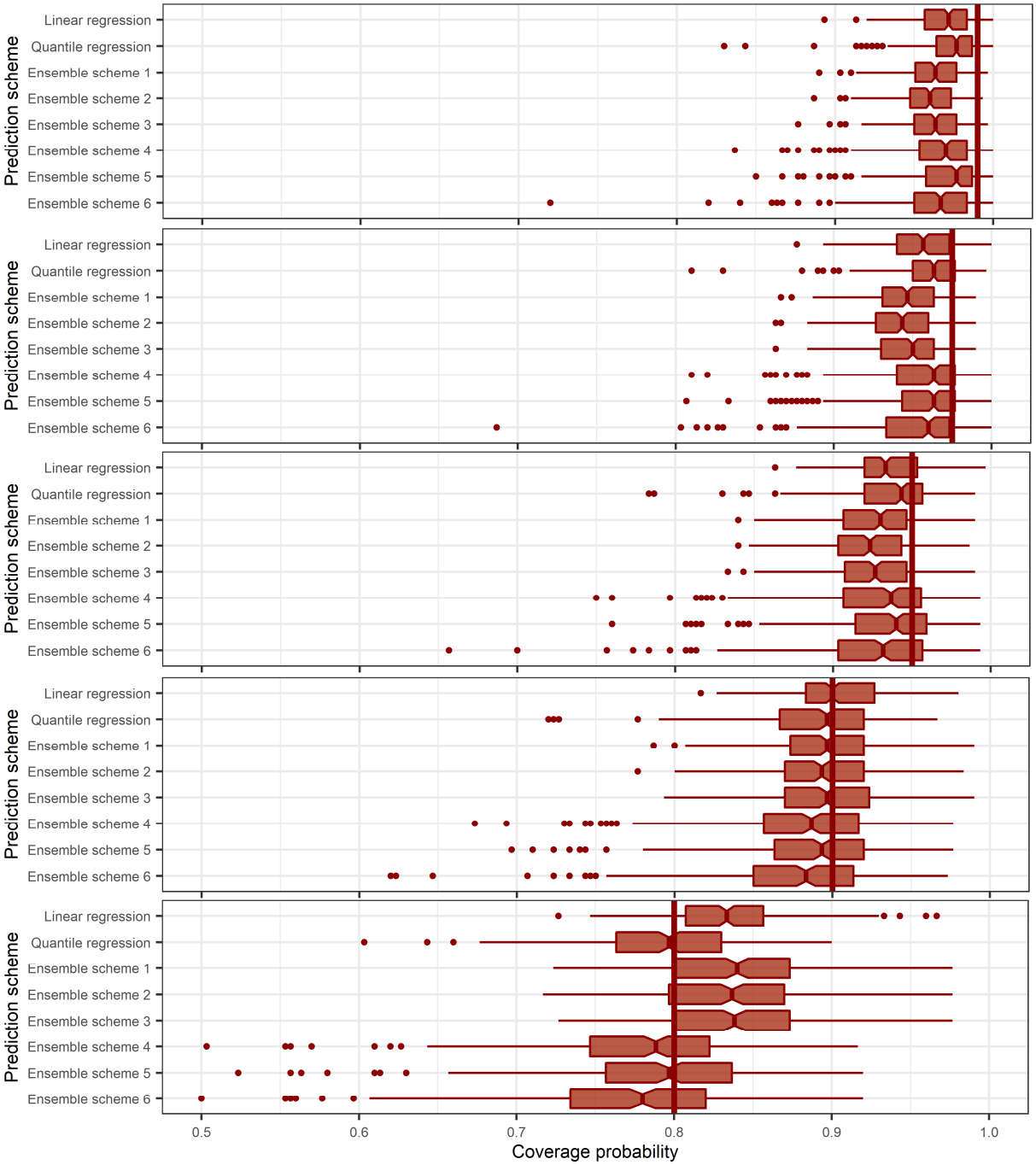

Figure 4. Coverage probabilities computed for the 99%, 97.5%, 95%, 90% and 80% prediction intervals (from top to bottom) delivered by the compared schemes for the period T_3 (years 1975–1999). Each boxplot summarizes 270 values. The optimal values are denoted with red thick vertical lines.

Table 2. Average coverage probabilities computed for the prediction intervals delivered by the compared schemes for the period T_3 (years 1975–1999). Each presented value summarizes 270 metric values.

Prediction scheme	99% prediction intervals	97.5% prediction intervals	95% prediction intervals	90% prediction intervals	80% prediction intervals
Linear regression	0.969	0.955	0.937	0.904	0.835
Quantile regression	0.973	0.961	0.936	0.889	0.793
Ensemble scheme 1	0.962	0.946	0.926	0.895	0.834
Ensemble scheme 2	0.959	0.943	0.923	0.892	0.834
Ensemble scheme 3	0.962	0.946	0.926	0.895	0.837
Ensemble scheme 4	0.965	0.953	0.928	0.881	0.781
Ensemble scheme 5	0.969	0.956	0.932	0.886	0.789
Ensemble scheme 6	0.961	0.948	0.923	0.874	0.773

While the average-case reliability of all prediction schemes is remarkably high (see Table 2), the performance of the prediction schemes in terms of coverage probabilities varies from catchment to catchment (see Figure 4). The observed differences in performance become larger, e.g., in terms of interquartile range of the formed datasets, as we move from the 99% to the 80% prediction intervals. Moreover, although differentiations are observed between prediction schemes, the overall performance of most schemes is rather of the same quality (in particular for the outer prediction intervals), with the quantile regression scheme and ensemble scheme 5 to be the best-performing, especially the former one.

The average widths, on the other hand, clearly favour the ensemble schemes over the basic schemes (see Figure 5), with ensemble schemes 4–6 providing sharper predictions than ensemble schemes 1–3. In terms of the same criterion, ensemble schemes from the former (latter) category exhibit remarkably close performance to each other. The same applies in terms of coverage probabilities. As already expected because of the large differences observed in the river discharge regimes of the examined catchments, the average widths of the prediction intervals may differ significantly from catchment to catchment. These differences become smaller, as we move from the outer to the inner prediction intervals, i.e., from the 99% to the 80% prediction intervals.

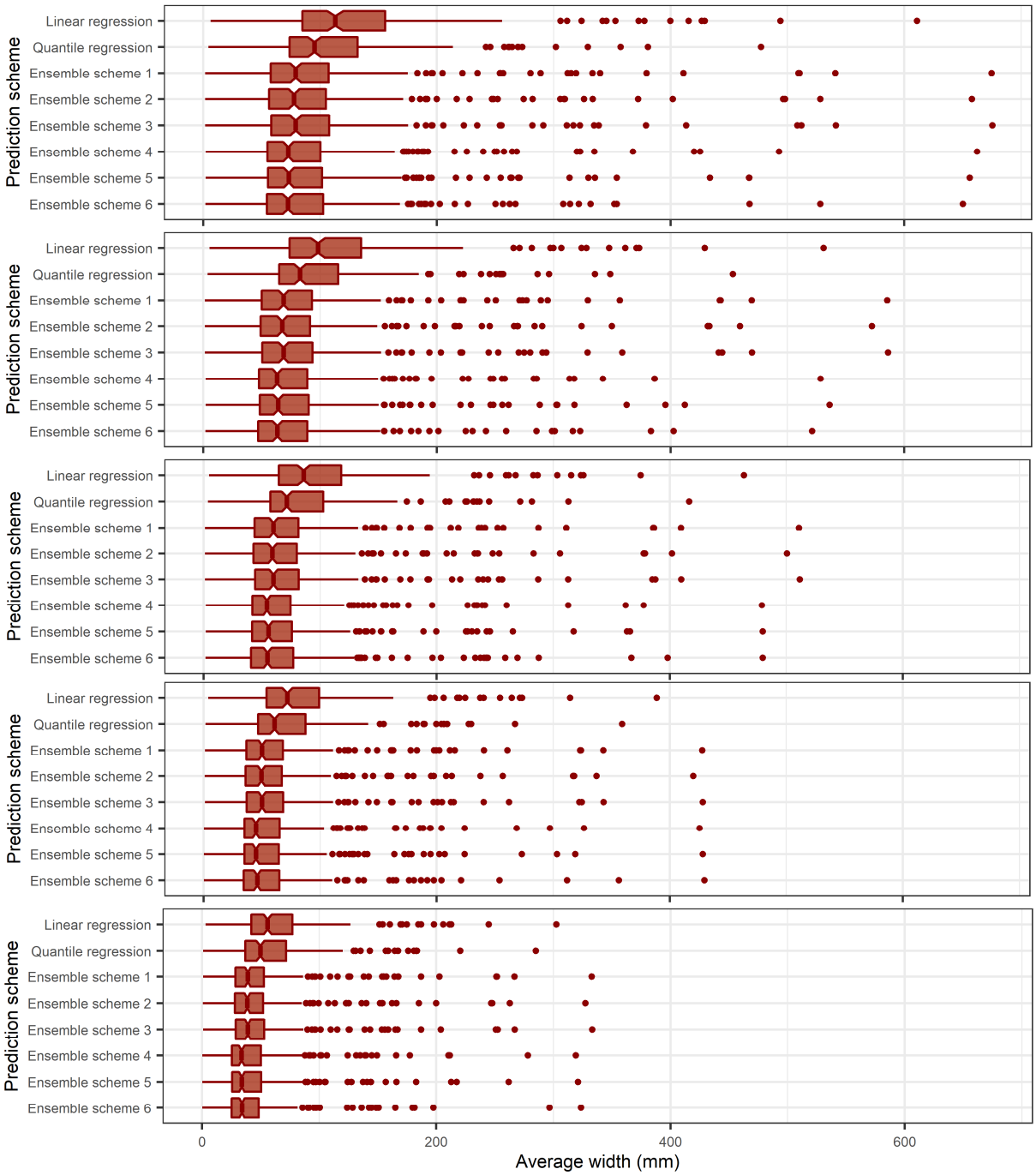

Figure 5. Average widths computed for the 99%, 97.5%, 95%, 90% and 80% prediction intervals (from top to bottom) delivered by the compared schemes for the period T_3 (years 1975–1999). Each boxplot summarizes 270 values.

The above-outlined information is objectively summarized in the average interval scores. The latter are collectively presented in Figure 6. The main information extracted from this figure is that (a) ensemble schemes 1–3, as well as ensemble schemes 4–6, exhibit very close performance to each other, (b) each ensemble scheme exhibits a better overall performance than its corresponding basic scheme, and (c) ensemble schemes 1–3 perform better than the quantile regression scheme for the 90% and 80%

prediction intervals. Observation (b) indicates that the herein adopted implementations of the working methodology have an advantage over the naïve implementations of the data-driven (or purely statistical) models. This advantage should be further investigated before any generalization is made; nevertheless, this additional investigation involving, for instance, utilization of more predictor variables, goes beyond the aim of the present study.

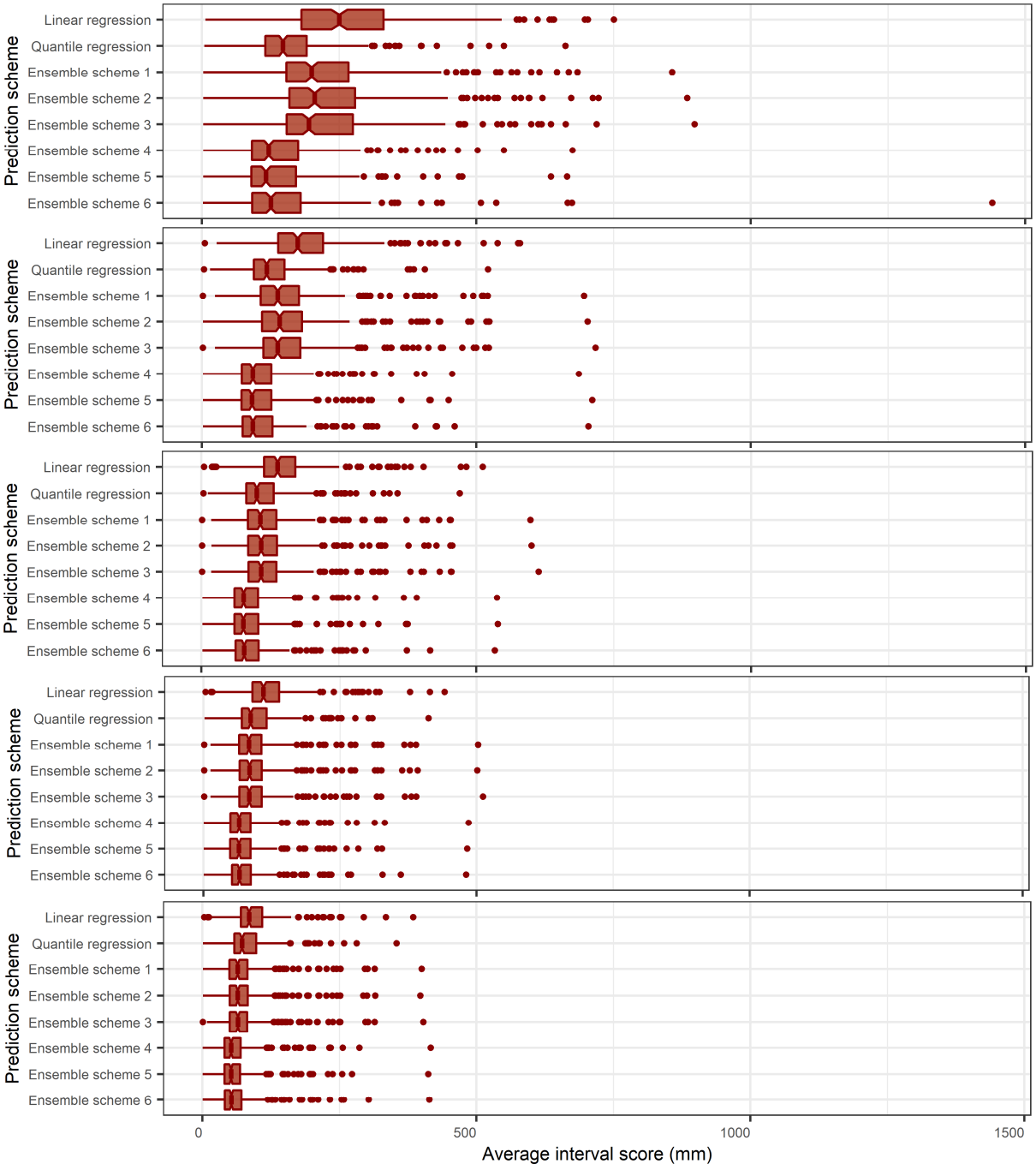

Figure 6. Average interval scores computed for the 99%, 97.5%, 95%, 90% and 80% prediction intervals (from top to bottom) delivered by the compared schemes for the period T_3 (years 1975–1999). Each boxplot summarizes 270 values.

We also note that both observations (a) and (b) are roughly expected already from the examination of Figures 4 and 5. By examining the aggregated average interval scores we additionally observe that the differences with respect to this metric are in average smaller for the inner prediction intervals than for the outer ones (as expected; see Section 2.3). Some small differences in the performance of ensembles schemes 1–3, favouring to a small extent ensemble schemes 1 and 3 over ensemble scheme 2, are mostly noticeable for the 99% and 97.5% prediction intervals. Similarly, ensemble scheme 5 seems to perform slightly better than ensemble scheme 4 for the same prediction intervals. It is also more effective than ensemble scheme 6 for all five prediction intervals.

To further inspect all differences, both the smaller and larger ones, in terms of rankings, the latter resulted for each catchment and for each examined prediction interval according to the computed average interval scores, we present Figures 7 and 8. The maps displayed in the former figure correspond to the upper side-by-side boxplots displayed on Figure 6, while allowing the examination of the rankings resulted both per catchment and per prediction scheme. From these maps we perceive that ensemble scheme 5 is ranked in a better average position than the remaining prediction schemes for the 99% prediction intervals, closely followed by ensemble schemes 4 and 6. Moreover, the quantile regression scheme is mostly ranked above the linear regression scheme and ensemble schemes 1–3. These schemes are mostly ranked in the last four positions. Importantly, there is not a fixed ranking position for any of the prediction schemes across the various catchments, while there are also some few catchments in which the four less competitive ones perform better than some the remaining. The quantile regression scheme is also ranked in the first three positions for a sufficient number of catchments. These latter observations provide us with a good reason to always perform large-scale benchmark experiments instead of (or alongside with) case studies.

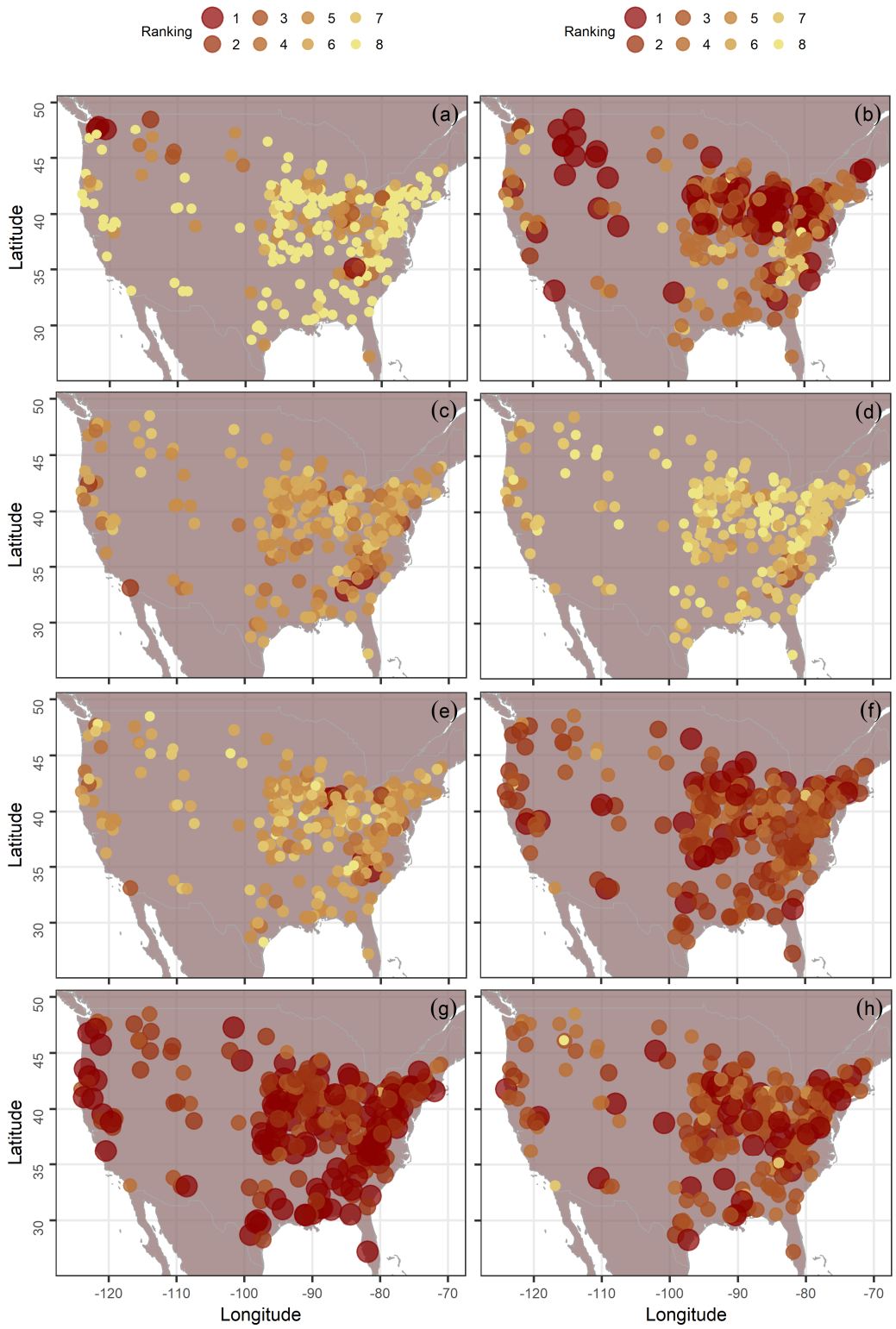

Figure 7. Rankings of (a) linear regression, (b) quantile regression and ensemble schemes (c–h) 1–6 according to the average interval scores computed for the 99% prediction intervals delivered for the period T_3 (years 1975–1999). The prediction schemes are ranked from best (1st) to worst (8th).

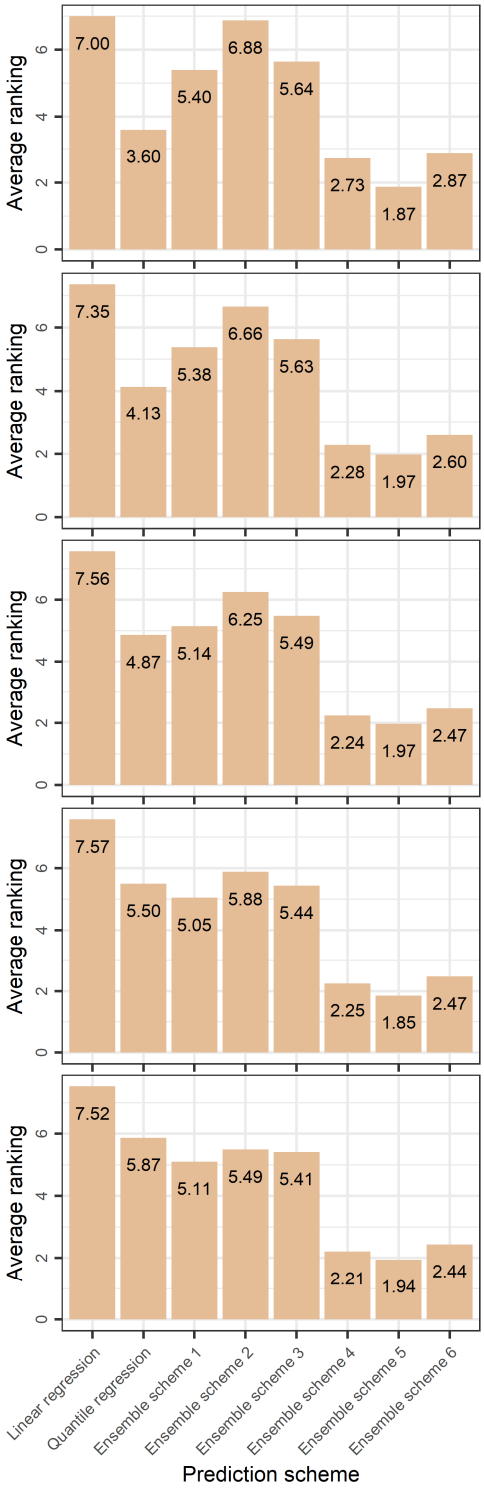

Figure 8. Average rankings of the prediction schemes according to the average interval scores computed for the 99%, 97.5%, 95%, 90% and 80% prediction intervals (from top to bottom) delivered by the compared schemes for the period T_3 (years 1975–1999). The prediction schemes are ranked from best (1st) to worst (8th). Each bar summarizes 270 values.

Overall, the image depicted in Figure 7 is rather neat when contrasted with its corresponding image in a similar visualization by Tyralis and Papacharalampous (2018); see Figure 4 therein. The latter study presents a large-scale comparison of point

prediction methods that are equivalent to each other in a long run; therefore, no pattern is observed in their performance when the latter is depicted in maps. The pattern clearly observed in Figure 7, favouring the quantile regression model over the linear regression one, is due to the suitability of the former algorithm for modelling heteroscedasticity. Thus, it is our knowledge on the examined problem and the difference in the appropriateness of the adopted methodologies that created this pattern rather than anything else.

As emphasized in Papacharalampous et al. (2019a), only our knowledge on the system could make a tangible difference in (predictive) modelling in a long run. In fact, the homoscedasticity assumption is known to be inefficient when made during the probabilistic modelling of hydrological variables, such as the monthly river discharge variables that are of interest herein (see the comments, e.g., in Schoups and Vrugt 2010; Montanari and Koutsoyiannis 2012; Evin et al. 2013, 2014). Therefore, more flexible algorithms not assuming homoscedasticity are a reasonable choice to be made in such cases, while the same algorithms do not offer anything in comparison with less flexible algorithms in modelling cases where the homoscedasticity assumption is reasonable; see also Papacharalampous et al. (2019b), in particular the results displayed in Tables 4 and 5 for an illustration-justification of this fact.

The greatest part of the ranking-related information extracted from Figure 7 applies as well to the remaining prediction intervals, while a summary of this information for the 99%, 97.5%, 95%, 90% and 80% prediction intervals, presented in Figure 8, provides additional observations. The latter effectively complement those obtained from Figure 6. In fact, for all prediction intervals ensemble scheme 5 exhibits the best average-case ranking, closely followed by ensemble schemes 4 and 6. Moreover, the quantile regression scheme exhibits a significantly better (comparable) average-case ranking than (with) ensemble schemes 1–3 for the 99% and 97.5% (95%, 90% and 80%) prediction intervals, while the linear regression scheme is the worst performing in terms of average rankings, as it could be expected already from Figure 6.

To obtain a more faithful image of the gain or loss in performance when using each prediction scheme over the remaining ones, in Figure 9 we present the side-by-side boxplots of the relative improvements in terms of average interval score with respect to the linear regression scheme, while in Figure 10 we present the respective information using the quantile regression scheme as a reference. The closeness in the performance of

ensembles schemes 1–3 is also perceivable by the examination of these figures. The same applies to the closeness in the performance of ensemble schemes 4–6. Nevertheless, the small differences favouring ensemble schemes 1 and 3 over ensemble scheme 2, and ensemble scheme 5 over ensemble schemes 4 and 6 are also highlighted. Additionally, we observe that the differences in the relative performance of a specific prediction scheme can be large, while there are cases in which the ensemble schemes are (far) worse than their respective basic schemes. However, the long-run image clearly favours the former over the latter, as already expected from the preceding visualizations.

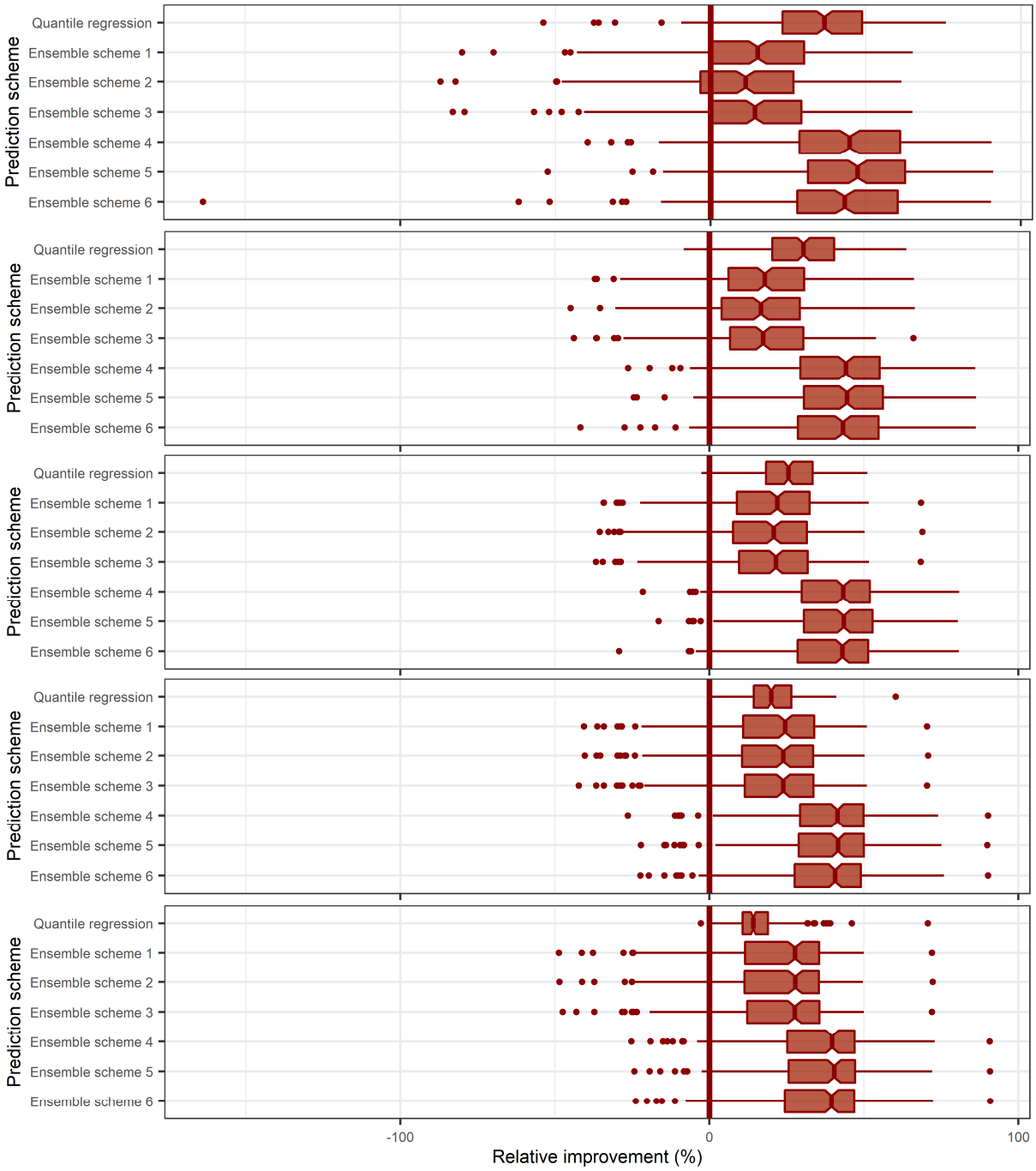

Figure 9. Relative improvements in terms of average interval score with respect to the linear regression scheme for the 99%, 97.5%, 95%, 90% and 80% prediction intervals (from top to bottom) delivered by the compared schemes for the period T_3 (years 1975–1999). Each boxplot summarizes 270 values. The reference values (zero values) are denoted with red thick vertical lines.

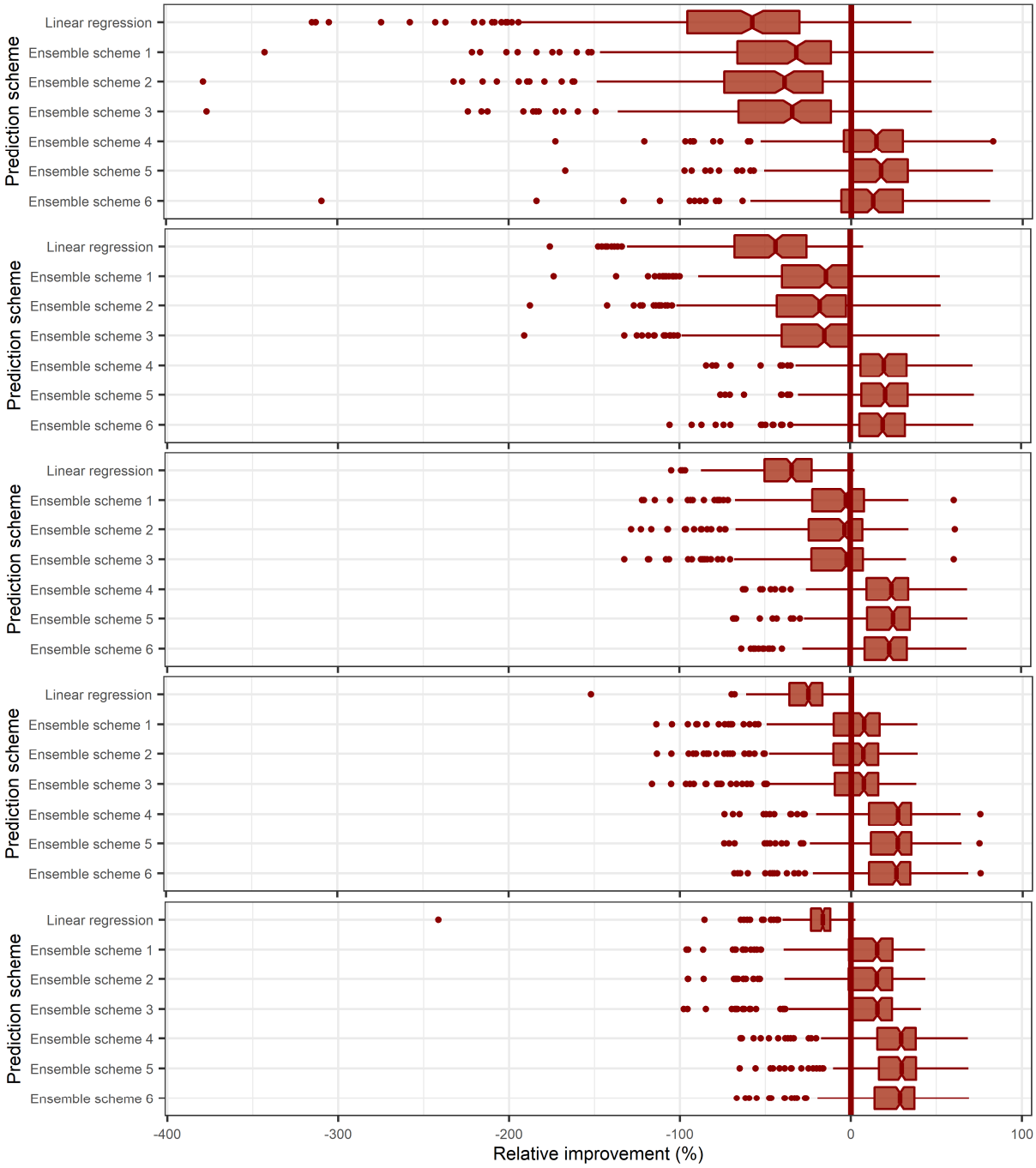

Figure 10. Relative improvements in terms of average interval score with respect to the quantile regression scheme for the 99%, 97.5%, 95%, 90% and 80% prediction intervals (from top to bottom) delivered by the compared schemes for the period T_3 (years 1975–1999). Each boxplot summarizes 270 values. The reference values (zero values) are denoted with red thick vertical lines.

We subsequently provide a numerical summary of the gain in performance when using specific schemes over others, as extracted from the real-world experiment of the study. In Figures 11 and 12 we present the average-case relative improvements in terms of average interval score with respect to the linear regression and the quantile regression schemes respectively. These two figures objectively summarize the

information presented in Figures 9 and 10, while they are particularly useful in assessing how small the differences between ensemble schemes 1−3, as well as between ensembles schemes 4−6, are; see also Figures S.1 and S.2 of the supplementary material (see Appendix D) for inspecting these differences in terms of median relative improvements. For the former category of ensemble schemes, we observe that the difference in the average-case improvements is at maximum 3.65%. The latter difference is computed for ensemble schemes 1 and 2 for the 99% prediction intervals, while it is smoothened to 1.94%, 1.07%, 0.48% and 0.13% for the 97.5%, 95%, 90% and 80% prediction intervals respectively. The average relative improvements when using ensemble scheme 1 instead of ensemble scheme 2 are 4.24%, 2.39%, 1.36%, 0.63% and 0.18% for the obtained 99%, 97.5%, 95%, 90% and 80% prediction intervals. The respective median improvements are 3.75%, 2.18%, 1.20%, 0.53% and 0.15%, while the cost in terms of computational time is about 12 min for all 270 catchments. Ensemble scheme 3 offers comparable profit in performance alongside with a 28-minute profit in terms of computational time compared to ensemble scheme 1.

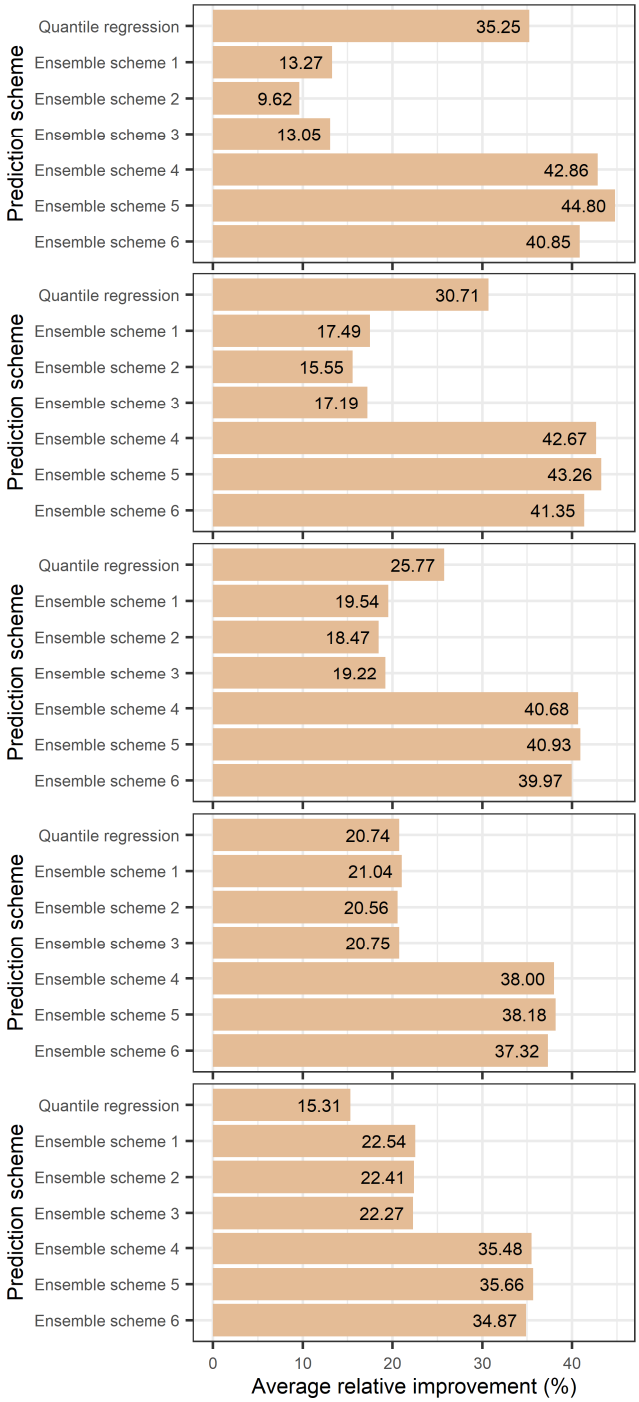

Figure 11. Average relative improvements in terms of average interval score with respect to the linear regression scheme for the 99%, 97.5%, 95%, 90% and 80% prediction intervals (from top to bottom) delivered by the compared schemes for the period T_3 (years 1975–1999). Each bar summarizes 270 values.

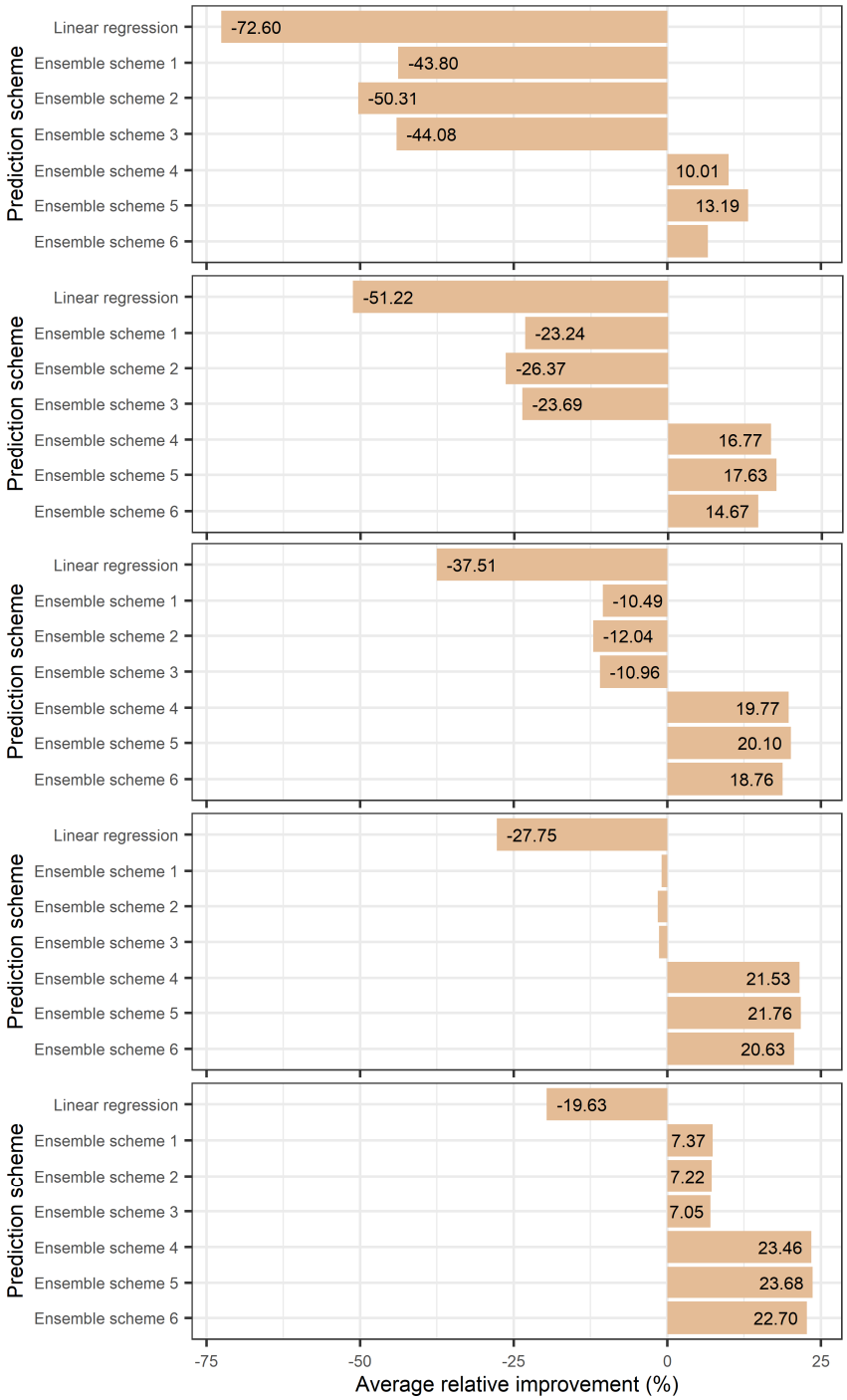

Figure 12. Average relative improvements in terms of average interval score with respect to the quantile regression scheme for the 99%, 97.5%, 95%, 90% and 80% prediction intervals (from top to bottom) delivered by the compared schemes for the period T_3 (years 1975–1999). Each bar summarizes 270 values.

Moreover, the mean (median) profit when using ensemble scheme 5 instead of ensemble scheme 4 is found to be 3.09%, 0.99%, 0.48%, 0.34% and 0.25% (2.07%, 0.54%, 0.32%, 0.27% and 0.18%) for the 99%, 97.5%, 95%, 90% and 80% prediction intervals respectively, while the concomitant cost in terms of computational time is about 36 min. The respective profit when using ensemble scheme 6 over ensemble

scheme 4 is about 12 min. Nonetheless, the use of the latter scheme instead of the former scheme offers an average (median) relative improvement equal to 2.23%, 1.77%, 1.11%, 1.00% and 0.85% (0.31%, 0.47%, 0.24%, 0.28% and 0.31%) for the 99%, 97.5%, 95%, 90% and 80% prediction intervals respectively. Moreover, the respective average (median) relative improvements provided by ensemble scheme 5 with respect to ensemble scheme 6 are 5.46%, 2.74%, 1.60%, 1.36%, 1.10% (3.39%, 1.44%, 0.73%, 0.57%, 0.45%). The gain in performance from the incorporation into the working methodology of the quantile regression model instead of the linear regression model can be summarized by the average-case (median) relative improvements in terms of average interval score provided when using ensemble scheme 5 instead of ensemble scheme 1. These are 37.00%, 31.62%, 26.82%, 22.10% and 17.22% (37.97%, 31.32%, 25.85%, 20.95% and 15.84%) for the 99%, 97.5%, 95%, 90% and 80% prediction intervals respectively.

3.2 Addressing aims 4–5 of the study

Two key properties of the working methodology, as identified in Papacharalampous et al. (2019b) based on the seminal work by Lichtendahl et al. (2013, Section 5), are its larger robustness in performance compared to basic two-stage post-processing methodologies and its ability to harness the wisdom of the crowd, both stemming from the concept of prediction averaging. These properties can also be considered as the result of an optimal exploitation of the possibilities offered by the MK blueprint methodology. The demonstration of these properties has only been made so far within toy examples, while it is still pending for rainfall-runoff problems. This section is devoted to empirically proving these two properties of the working methodology using the results of the herein conducted real-world experiment, i.e., to addressing aims 4–5 of the study. These aims are of particular importance in justifying the conceptualization and rationale behind the working methodology.

In Figure 13 we present the relative improvements when using the output of ensemble scheme 5, i.e., the average of 600 quantile predictions, instead of separately using each of them (i.e., the relative improvements $\{RI_{OUT,IN_i}, i = 1, ..., 600\}$, defined with Equation 6, for ensemble scheme 5), computed for all catchments and for all prediction intervals. We observe that these relative improvements are approximately symmetric around zero, in average slightly higher than zero. Specifically, the average relative

improvements corresponding to Figure 13 are found to be equal to 0.82%, 0.83%, 0.74%, 0.70% and 0.71% for the 99%, 97.5%, 95%, 90% and 80% prediction intervals respectively (see Table S.1). The interpretation of this outcome is straightforward, while indicating an advantage in terms of robustness of the working methodology over basic two-stage post-processing methodologies using a single probabilistic prediction.

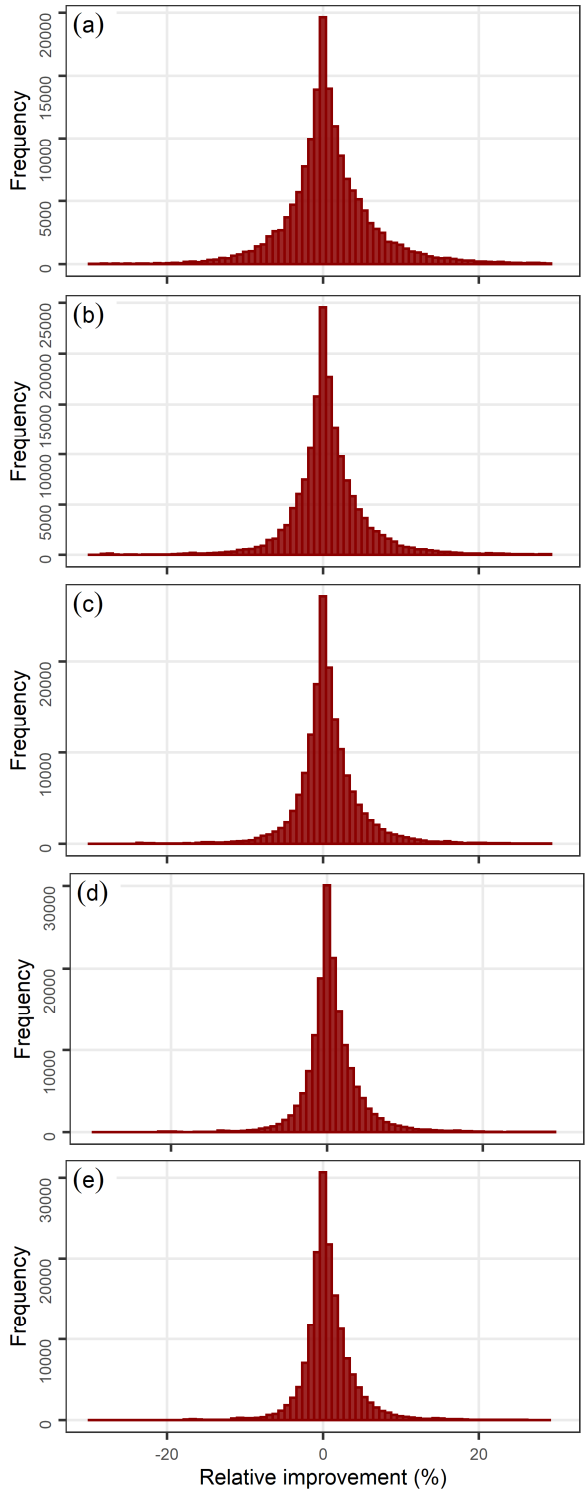

Figure 13. Relative improvements $\{RI_{OUT,IN_i}, i = 1, ..., 600\}$ (defined with Equation 6) for ensemble scheme 5. The relative improvements are computed for all catchments, and for the (a) 99%, (b) 97.5%, (c) 95%, (d) 90% and (e) 80% prediction intervals obtained for the period T_3 (years 1975–1999). The horizontal axis has been truncated at −30% and 30%. Each histogram summarizes 270 × 600 = 162 000 values.

In fact, while approximately half of the probabilistic predictions score better (or worse) than the finally delivered by the working methodology probabilistic prediction, there is no way to know in advance which hydrological model's parameters will lead in

better average interval score in the period T_3. While this lack of knowledge could significantly affect (in terms of performance) the delivered probabilistic prediction for a basic two-stage post-processing methodology, this effect is largely reduced by the working methodology.

Moreover, by comparing the degree of spread in the five histograms displayed in Figure 13, we also perceive that the degree of the offered stabilization in performance seems to become larger as we move from the inner prediction intervals to the more outer ones. Nevertheless, even for the 80% prediction intervals the provided stabilization is significant.

Furthermore, in Figure 14 we present the relative differences between the average interval score of the output of ensemble scheme 5 and the average of the average interval scores of each of the combined (for obtaining this output) individual predictions, the latter used as reference for the former (i.e., the relative differences $RD_{OUT,AAIS_{IN}}$, defined with Equation 8, for ensemble scheme 5), computed for all catchments and for all prediction intervals. Importantly, all computed relative differences are positive (or approximately zero) with no exception; therefore, the average of quantile predictions scores no worse than the average score of the combined individual predictions, i.e., the working methodology harnesses the wisdom of the crowd in terms of average interval score when applied for solving monthly rainfall-runoff problems (see also Lichtendahl et al. 2013, Section 5). The average relative differences corresponding to Figure 14 are 1.30%, 1.12%, 0.94%, 0.85% and 0.84% for the 99%, 97.5%, 95%, 90% and 80% prediction intervals respectively (see Table S.2).

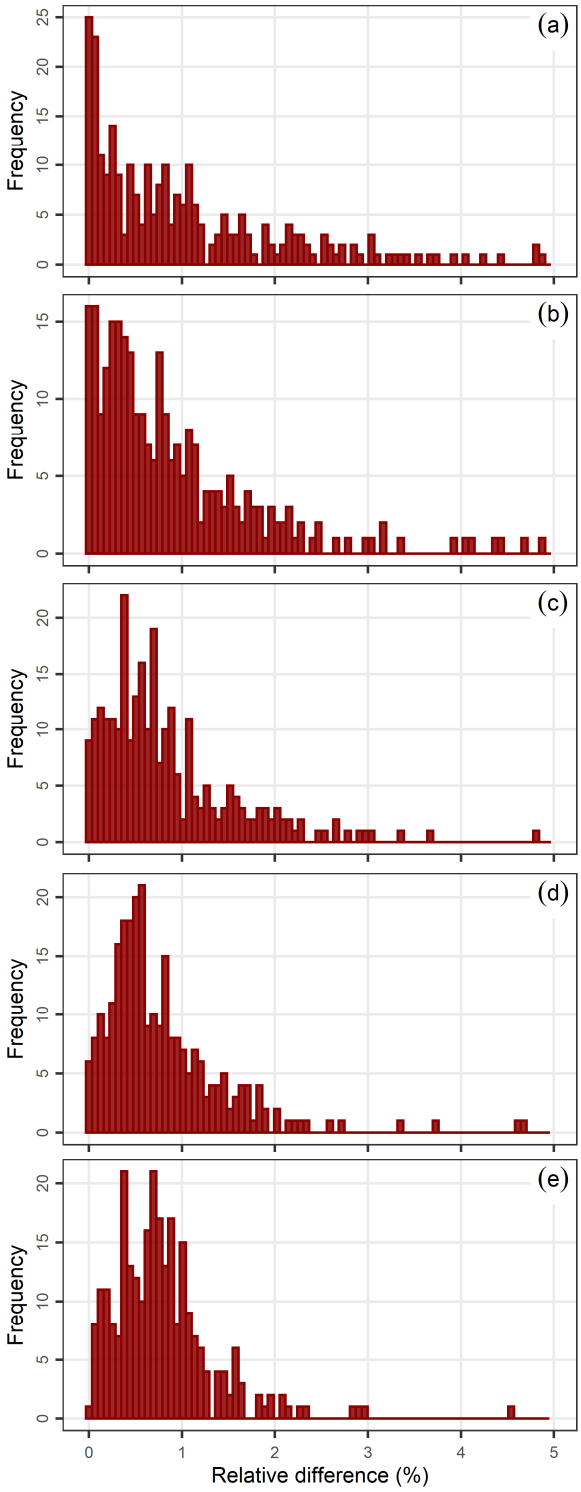

Figure 14. Relative differences $RD_{OUT,AAIS_{IN}}$ (defined with Equation 8) for ensemble scheme 5. The relative differences are computed for all catchments, and for the (a) 99%, (b) 97.5%, (c) 95%, (d) 90% and (e) 80% prediction intervals obtained for the period T_3 (years 1975–1999). The horizontal axis has been truncated at 5%. Each histogram summarizes 270 values.

Analogous observations are extracted from analogous investigations for all remaining ensemble schemes (see Figures S.3–S.12 and Tables S.1–S.2 of the supplementary material). In summary, the relative improvements when using the

output of an ensemble scheme, i.e., the average of 600 quantile predictions, instead of separately using each of these predictions range from −327.10% to 91.42%. The average of these relative improvements ranges between 0.13% and 1.13%. Similarly, the average relative differences favouring the average interval score computed for the output of an ensemble scheme over the average of the average interval scores computed for each of the combined (for obtaining this output) individual predictions range between 0.19% and 1.83%. The average relative improvement (difference) is in general larger for the outer prediction intervals than for the inner ones, while its magnitude also depends on the ensemble scheme.

As also emphasized in Papacharalampous et al. (2019b), the overall trade-off to be considered when someone has to choose between the working methodology and a basic two-stage post-processing methodology allowing the utilization of the same type of flexible error models (see e.g., López López et al. 2014; Dogulu et al. 2015; Papacharalampous et al. 2019d) is the one between (a) the larger robustness in performance offered by the former methodology (demonstrated in Figures 13, S.3, S.5, S.7, S.9 and S.11, and Table S.1) and the ability of this methodology to harness the wisdom of the crowd (empirically proven based on Figures 14, S.4, S.6, S.8, S.10 and S.12, and Table S.2), and (b) the significantly less computational requirements of the latter methodologies.

4. Concluding remarks

We have validated the probabilistic hydrological modelling methodology proposed in Papacharalampous et al. (2019b). This methodology adopts key concepts from the ensemble post-processing methodology by Montanari and Koutsoyiannis (2012), while also relying on the concept of probabilistic prediction combination from the forecasting field. It applies a single hydrological model using a large number of different parameter values to generate the same number of "sister predictions". The parameters of the hydrological model can be obtained by using either Bayesian calibration schemes or informal calibration schemes (see the related investigations in Appendix E). Therefore, this methodology does not have any particular relationship with Bayesian methods by construction, as it also applies to its precursor. A statistical learning (or machine learning) regression model that is suitable for predicting quantiles (see e.g., the models exploited in Papacharalampous et al. 2019d) is then used to obtain information about

the hydrological model's error. This information is used to convert the sister predictions into probabilistic predictions, which are finally combined in simple fashion to obtain the output probabilistic predictions. The assessed methodology is subdivided into three alternative variants, which differ only in the training of the regression model.

We have conducted a large-sample real-world experiment at monthly timescale, set up using complete 50-year daily information for 270 catchments in the United States. Aiming to increase the understanding in probabilistic hydrological modelling, we have insisted on interpretability and benchmarking within all conducted tests. We have used the parsimonious GR2M hydrological model and two (largely) interpretable regression models, specifically the linear regression and the quantile regression ones, to implement six ensemble schemes, all of them based on the assessed methodology. Those ensemble schemes implemented using the linear model (three in number) have been used as benchmarks for the remaining schemes (also three in number). Those ensemble schemes using the same regression model rely on different variants of the assessed methodology. The performance of the ensemble schemes has been assessed by computing the coverage probabilities, average widths and average interval scores of the obtained interval predictions, and by also benchmarking their results using naïve probabilistic data-driven models.

The obtained numerical results (metric values computed for 4 870 800 interval predictions) suggest the usefulness of the assessed methodology in obtaining probabilistic predictions of hydrological quantities. The best-performing variant, offering a mean relative improvement up to 5.46% with respect to its alternative variants, when implemented using the quantile regression model, is variant 2. This variant trains the regression model on a single large dataset formed by using information from all sister predictions. The average-case relevant improvements when using the quantile regression model instead of the linear regression one range up to about 37% in terms of average interval score. This latter numerical result should be appraised on the basis that only the former of these models can model heteroscedasticity. The homoscedasticity assumption is often made in the literature when modelling the hydrological model's error.

Finally, we have demonstrated the increased robustness of the assessed methodology with respect to the combined (by this methodology) individual predictors and, by extension, to basic two-stage post-processing methodologies. The ability to

"harness the wisdom of the crowd" has also been empirically proven. The quantile predictions obtained by all ensemble predictors are found to score no worse –usually better– than the average of the individual scores of the combined individual predictions in terms of average interval score. This outcome is in line with demonstrations for stylized cases by Lichtendahl et al. (2013). The computed relative differences favour the former quantity over the latter up to about 37%, while their mean values range between 0.19% and 1.83%, depending both on the prediction interval and the variant of the assessed methodology. For the best-performing ensemble scheme the respective average relative differences are around 1%. Overall, the robustness and the ability to harness the wisdom of the crowd are identified as two key properties of the working methodology.

Appendix A Background methodological considerations

In this appendix, we summarize in terms of advantages and disadvantages some technical and theoretical considerations that currently guide the selection between Bayesian and two-stage post-processing methodologies for uncertainty assessment in the field. This summary is mainly presented through Tables A.1 and A.2. Moreover, in Table A.3 we list the advantages and disadvantages offered by statistical learning (or machine learning) quantile regression algorithms, since these algorithms serve as error models within the working methodology.

Table A.1. Advantages and disadvantages of Bayesian hydrological post-processing methodologies (see also Evin et al. 2014). These post-processing methodologies jointly infer (within a Bayesian framework) the parameters of the hydrological and error models by using the entire historical dataset.

Advantages	○ If their assumptions are proper, they produce optimal probabilistic predictions by theory. This could be possible in principle, since the hydrological literature presents generalized findings on the distributions of hydrological variables with increasing frequency and reliability. ○ They can largely facilitate interpretability in modelling, since they allow the inspection of the impact of their assumptions on both parameter and predictive uncertainty. ○ Their performance depends less on the length of the historical dataset than the performance of two-stage post-processing methodologies (see Table A.2), since their fitting does not require sample splitting.
Disadvantages	○ Their predictive performance largely depends on the appropriateness of their assumptions. ○ They might get over-parameterized in an effort to ensure the adoption of proper assumptions. ○ Their use is accompanied by computational limitations.

Table A.2. Advantages and disadvantages of two-stage hydrological post-processing methodologies (see also Evin et al. 2014; Papacharalampous et al. 2019d, Section 5.2.2). These post-processing methodologies estimate their error models conditional on the predictions provided by their hydrological models. The latter have been calibrated by using an independent segment of the historical dataset.

Advantages	○ They can be nearly assumption-free (i.e., their performance does not necessarily depend on the appropriateness of assumptions) when implemented with flexible machine learning quantile regression algorithms as error models. The advantages of these algorithms are listed independently in Table A.3. ○ Computational requirements and limitations are mostly few in their case. Therefore, their automation and application to big datasets is feasible. This is one of the main reasons why two-stage hydrological post-processing is popular in forecasting applications. This popularity is emphasized e.g., by Evin et al. (2014). ○ In light of the two points above, their performance can be maximized by adopting algorithmic strategies and well-established guidelines from the machine learning literature (see e.g., the experiment presented herein). The role of big datasets for achieving optimal modelling solutions under this new-era approach is emphasized e.g., in Tyralis et al. (2019b).
Disadvantages	○ They largely lack interpretability by perception. Interactions between the hydrological model parameters and the trained version of the error model are ignored; therefore, their hydrological model parameter estimates are only auxiliary to predictive uncertainty quantification and cannot be used in any case for understanding parameter uncertainty. ○ Their performance depends more on the length of the historical dataset than the performance of Bayesian post-processing methodologies (see Table A.1), since their fitting requires sample splitting. ○ The adoption of flexible machine learning quantile regression algorithms as error models has an additional cost in terms of interpretability and further increases the large-sample requirements (see the disadvantages of Table A.3). These requirements are revealed and discussed e.g., in Papacharalampous et al. (2019b, Appendix D).

Table A.3. Advantages and disadvantages of statistical learning (or machine learning) quantile regression algorithms (see also Waldmann 2018; Papacharalampous et al. 2019d, Sections 2.3.1, 5.2.2). Quantile regression algorithms issue quantile predictions instead of PDF predictions.

Advantages	○ They are ideal when the conditional distribution of the dependent variable is not known or is hard to deduce. ○ They model heteroscedasticity by perception and construction. ○ In light of the above point, they are also straightforward to apply, as they do not need to be fitted separately for each season (or month), in contrast to distribution-based modelling approaches (e.g., conditional-distribution models). ○ They are robust with respect to outliers in the observations of the dependent variable. ○ They are available in open source and mostly optimally programmed.
Disadvantages	○ They are trained separately for each quantile probability; therefore, the more the quantiles (or prediction intervals) we are interested in issuing, the more computationally costly the training process. ○ Quantile crossing is possible. ○ Parameter estimation is harder than in standard regression. ○ Their performance depends to some extent on the sample size. ○ They lack interpretability. Only their linear variant, i.e., the quantile regression model implemented herein, offers interpretability to some extent.

Appendix B Statistical software information

The analyses and visualizations have been performed in `R` Programming Language (`R` Core Team 2019). We have used the following contributed `R` packages: `airGR` (Coron et al. 2017, 2019), `bestNormalize` (Peterson 2017, 2019), `coda` (Plummer et al. 2006; 2019), `data.table` (Dowle and Srinivasan 2019), `devtools` (Wickham et al. 2019c), `dplyr` (Wickham et al. 2019b), `FME` (Soetaert and Petzoldt 2010, 2016), `gdata` (Warnes et al. 2017), `ggplot2` (Wickham 2016a; Wickham et al. 2019a), `ggridges` (Wilke 2018), `hddtools` (Vitolo 2017, 2018), `knitr` (Xie 2014, 2015, 2019), `maps` (Brownrigg et al. 2018), `matrixStats` (Bengtsson 2018), `plyr` (Wickham 2011, 2016b), `quantreg` (Koenker 2019), `readr` (Wickham et al. 2018), `reshape` (Wickham 2007, 2018), `rmarkdown` (Allaire et al. 2019), `tidyr` (Wickham and Henry 2019) and `zoo` (Zeileis and Grothendieck 2005; Zeileis et al. 2019). We have also followed procedures described in the contributed vignettes of the `airGR` `R` package (https://cran.r-project.org/web/packages/airGR/vignettes).

Appendix C Working methodology

This appendix is largely adapted from Papacharalampous et al. (2019b). It aims at

summarizing the working methodology. For this summary, we first define the time period $T = \{1, \ldots, (n_1+n_2+n_3)\}$, and its three distinct sub-periods $T_1 = \{1, \ldots, n_1\}$, $T_2 = \{(n_1+1), \ldots, (n_1+n_2)\}$ and $T_3 = \{(n_1+n_2+1), \ldots, (n_1+n_2+n_3)\}$. We also define the sister model realizations as variants of a single hydrological model, each using different parameter values. The latter are obtained by calibrating the hydrological model in the period T_1. The calibration could be made by using either Bayesian schemes (e.g., Markov Chain Monte Carlo simulation sampling; see e.g., the procedures described in Section 2.2.4) or informal calibration schemes (see e.g., the procedures described in Appendix E). Let us assume that we obtain m sister model realizations, where m is adequately large. Each sister model realization is then applied in the period $\{T_2, T_3\}$. The m resulted sister predictions also extend in the period $\{T_2, T_3\}$. We subsequently compute the sister model realizations' errors in the period T_2 by using the sister predictions alongside with their corresponding target values.

Information about the sister model realizations' error is then obtained by training a statistical learning regression model that is suitable for predicting quantiles (hereafter referred to as "error model"; see e.g., the error models exploited in Papacharalampous et al. 2019d) in the period T_2. In particular, we regress the sister model realizations' error at time t (response variable) on selected predictor variables (e.g., the sister prediction at time t). For each sister prediction extending in the period T_3, we (a) predict a set of quantiles (with selected probabilities) of the sister model realization's errors using the information obtained at the preceding step, and (b) transform these predictive quantiles to auxiliary predictive quantiles of the hydrological process of interest (by subtracting them from their corresponding sister prediction). Finally, at each time $t \in T_3$ we group the auxiliary predictive quantiles of the hydrological process of interest based on their corresponding probability (e.g., probability 0.95) to average them over each group. The resulted time series are the output quantile predictions.

The basic steps adopted within the working methodology are also summarized in Figure B.1.

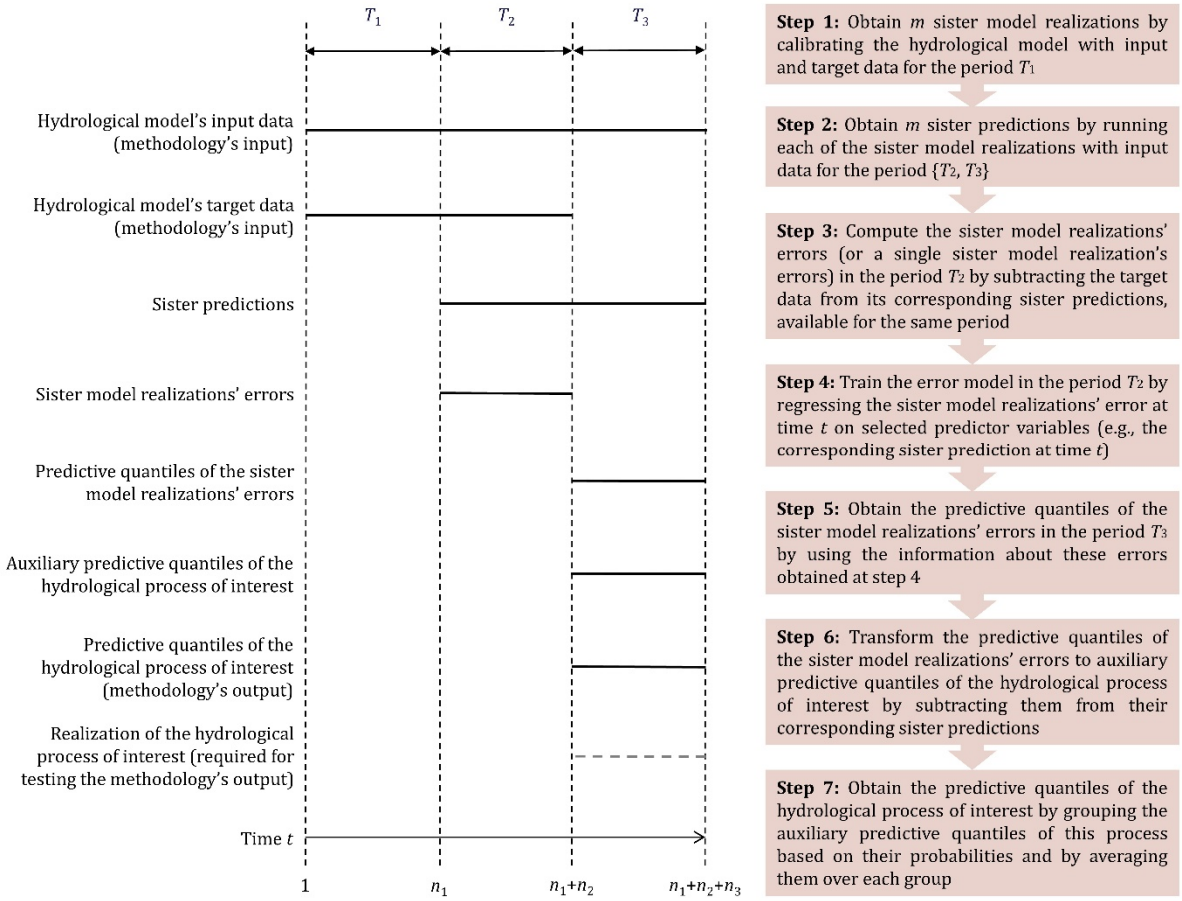

Figure B.1. Schematic summarizing the working methodology (reproduced from Papacharalampous et al. 2019b). The sister model realizations are defined as variants of a single hydrological model, each using different parameter values. The latter can either be drawn from the respective simulated posterior distribution of model parameters or can be obtained by using informal calibration schemes. Each sister model realization is used for obtaining a single point prediction, referred to as "sister prediction". The number of sister model realizations *m* should be adequately large. The realization of the hydrological process of interest, considered unknown at the time of the prediction, is denoted with a light grey dashed line.

The working methodology is subdivided into three alternative variants. These variants differ in the error model's training only. Specifically:

- Variant 1 trains the error model *m* times, each time on a different dataset formed by using a different sister prediction;
- Variant 2 trains the error model on a single dataset formed by using all sister predictions;
- Variant 3 also trains the regression model once; however, the training here is made on a dataset formed by using one randomly selected sister prediction.

We note that the three variants reduce to the same method in the case that a single point hydrological prediction is generated. In this case, the working methodology would fall

into the category of basic two-stage post-processing methodologies using regression models.

Appendix D Supplementary material

The supplementary material to this article is available in Papacharalampous et al. (2019c). This material includes Figures S.1–S.12, and Tables S.1 and S.2. The latter are extracted from the large-scale investigations presented in Section 3.

Appendix E Additional investigations

To investigate the possibility of using informal calibration schemes instead of Bayesian schemes for obtaining a large number of hydrological model's parameters within the working methodology, in this appendix we repeat the large-sample experiment of the study (only for the ensemble schemes) by using different parameter values for the hydrological model. Specifically, for each catchment we retain the first 200 parameter values from each simulated chain (see Section 2.2.4) that have not converged to the posterior distribution of the parameters, instead of the last 200 values that were previously retained (for the application presented in Section 3). Hereafter, let us refer to the calibration scheme adopted for obtaining the parameters of the hydrological model in the original large-sample experiment of the study (presented in Section 3) and the calibration scheme that is adopted in this appendix as "Bayesian calibration scheme" and "informal calibration scheme" respectively. The remaining components of the ensemble schemes are retained as detailed in Section 2.2.

Once we have obtained the interval predictions, we compute their interval scores and the relative improvements provided in terms of average interval score by the informal calibration scheme with respect to the Bayesian calibration scheme, when both these schemes are exploited as components of ensemble schemes 1–6. The computations are made as detailed in Section 2.3, while the related information is presented in Figure D.1. We mainly observe that (a) the relative improvements can be either positive or negative, and (b) the results favour the Bayesian calibration scheme to some extent, mostly due to outliers. These outliers may become fewer with increasing the length of the period T_2. To objectively summarize the derived information, we also compute the mean and median relative improvements in terms of the same score. These are presented in Figures D.2 and D.3 respectively.

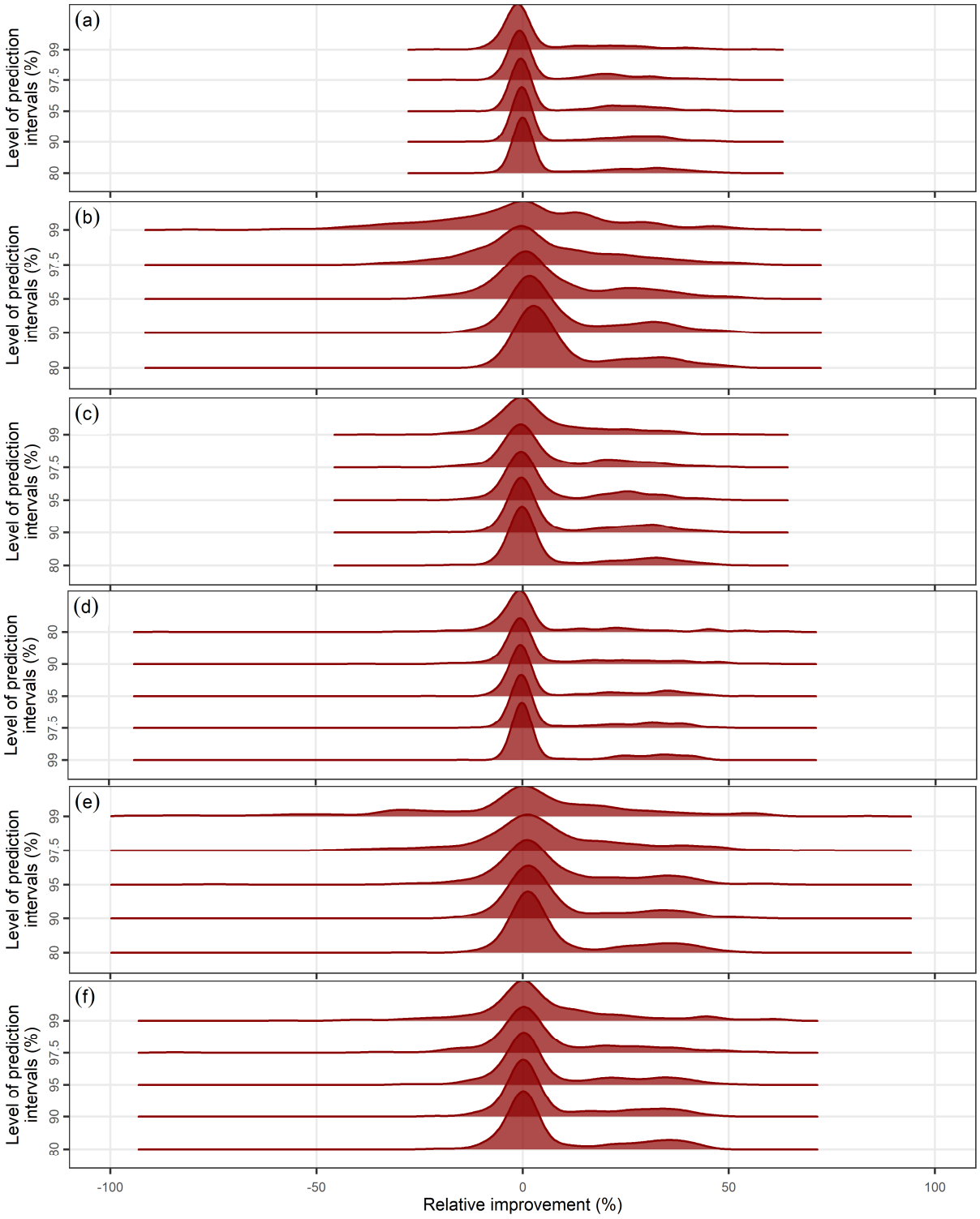

Figure D.1. Densities of the relative improvements in terms of average interval score provided by the Bayesian calibration scheme with respect to the informal calibration scheme, when both these schemes are used as components of (a–f) ensemble schemes 1–6. The latter are implemented with their remaining components and parameters set common. The horizontal axis has been truncated at −100% and 100%. Each density summarizes 270 values.

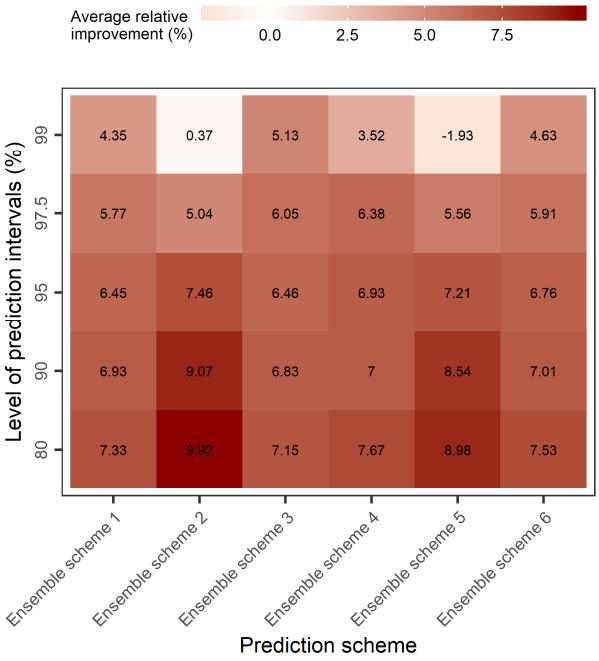

Figure D.2. Average relative improvements in terms of average interval score provided by the Bayesian calibration scheme with respect to the informal calibration scheme, when both these schemes are used as components of ensemble schemes 1–6. The latter are implemented with their remaining components and parameters set common. The legend limits are common for Figures D.2 and D.3. Each presented value summarizes 270 values.

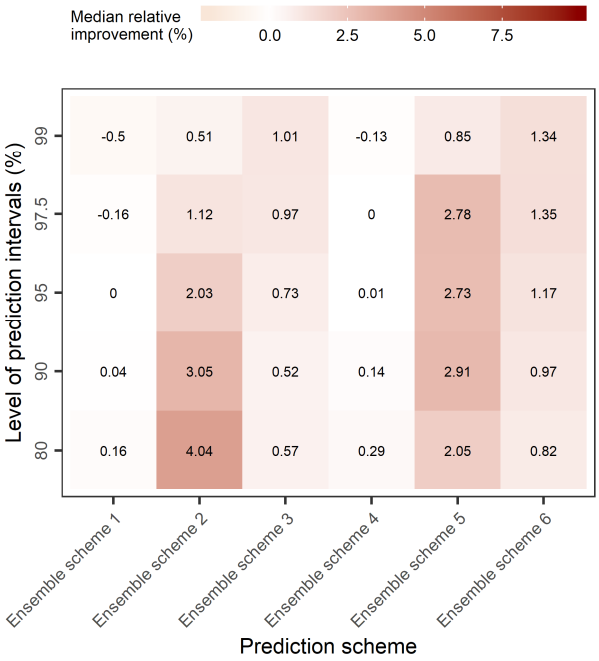

Figure D.3. Median relative improvements in terms of average interval score provided by the Bayesian calibration scheme with respect to the informal calibration scheme, when both these schemes are used as components of ensemble schemes 1–6. The latter are implemented with their remaining components and parameters set common. The legend limits are common for Figures D.2 and D.3. Each presented value summarizes 270 values.

Appendix F Additional remarks

We have extensively explored through benchmark tests the modelling possibilities provided by the working methodology, when this methodology is applied for solving monthly rainfall-runoff problems using the quantile regression model as error model. Our benchmark experiment is of large-scale; nevertheless, it could not highlight all aspects of the working methodology. For exploiting this methodology in an optimal way, the following key adjustments to its components and parameters could be made:

- The historical dataset can be divided in various ways, i.e., different proportions of the available information could be devoted to hydrological model calibration and error model training. This adjustment could be made to maximize predictive performance by exploiting evidence extracted from properly designed large-sample investigations. It could also be made for reducing the computational requirements, also depending on our choices on the remaining components and parameters. Applications to hundreds of catchments at timescales finer than the monthly one may require achieving a balance between predictive performance and computational requirements (when our computational resources are limited).
- Any hydrological model (e.g., a process-based hydrological model of our preference) can be selected. Predictive performance improvements may be achieved by selecting one hydrological model over another or by adopting multi-model approaches (as proposed in Vrugt 2018, 2019, yet with the interest being in producing and combining quantile predictions instead of PDF predictions), thereby extending the working methodology, as suggested by Montanari and Koutsoyiannis (2012) for the original blueprint. Properly designed large-sample investigations could effectively guide our related choices.
- The parameters of the hydrological model can be obtained by using a large variety of calibration schemes, including informal calibration schemes. (Note that random selection of the parameters, i.e., no period T_1, could also be an option). This point may be particular important for reducing the computational requirements. In Appendix E, we present large-sample investigations (on the monthly rainfall-runoff data exploited in the study) focusing on the comparison between Bayesian and informal calibration schemes for obtaining a large number of hydrological model parameters within the working methodology.

- The number of sister predictions can be selected based on the available computational resources. Nonetheless, the larger this number the larger the advantage of the methodology in terms of robustness (compared to basic two-stage post-processing methodologies). Properly designed benchmark experiments could also focus on optimizing this parameter of the working methodology (separately for the various timescales).
- Any statistical learning regression model that is suitable for predicting quantiles (e.g., the error models exploited in Papacharalampous et al. 2019d) can be selected as error model. This point may be particularly important for maximizing predictive performance (see also the key remarks in Section 4).
- Any set of predictor variables (e.g., the hydrological model predictions at times t, $t-1$, $t-2$, etc.) can be used in the application of the error model. This point may be important for maximizing predictive performance for timescales finer than the monthly one (see e.g., the findings in Papacharalampous et al. 2019d).
- All the above adjustments and modelling choices can be made separately for each of the three variants and for each level of prediction interval (or level of predictive quantile).

Acknowledgements: We sincerely thank the Editor, the Associate Editor, Dr Elena Volpi and an anonymous referee for their constructive reviews, which helped us to significantly improve the paper.

Declarations of interest: The authors declare no conflict of interest.

Funding information: The research work of Georgia Papacharalampous was supported by the Hellenic Foundation for Research and Innovation (HFRI) under the HFRI PhD Fellowship grant (Fellowship Number: 1388).

References

Abbas SA, Xuan Y (2019) Development of a new quantile-based method for the assessment of regional water resources in a highly-regulated river basin. Water Resources Management 33(8):3187–3210. https://doi.org/10.1007/s11269-019-02290-z

Abrahart RJ, See LM, Dawson CW (2008) Neural network hydroinformatics: Maintaining scientific rigour. In: Abrahart RJ, See LM, Solomatine DP (eds) Practical Hydroinformatics. Springer-Verlag Berlin Heidelberg, pp 33–47. https://doi.org/10.1007/978-3-540-79881-1_3

Allaire JJ, Xie Y, McPherson J, Luraschi J, Ushey K, Atkins A, Wickham H, Cheng J, Chang W, Iannone R (2019) `rmarkdown`: Dynamic Documents for `R`. `R` package version 1.14. https://CRAN.R-project.org/package=rmarkdown

Bengtsson H (2018) `matrixStats`: Functions that Apply to Rows and Columns of Matrices (and to Vectors). `R` package version 0.54.0. https://CRAN.R-project.org/package=matrixStats

Beven KJ (1993) Prophecy, reality and uncertainty in distributed hydrological modelling. Advances in Water Resources 16(1):41–51. https://doi.org/10.1016/0309-1708(93)90028-E

Beven KJ (2000) Uniqueness of place and process representations in hydrological modelling. Hydrology and Earth System Sciences 4:203–213. https://doi.org/10.5194/hess-4-203-2000

Beven KJ (2001) How far can we go in distributed hydrological modelling?. Hydrology and Earth System Sciences 5:1–12. https://doi.org/10.5194/hess-5-1-2001

Beven KJ (2006) A manifesto for the equifinality thesis. Journal of Hydrology 320(1–2):18–36. https://doi.org/10.1016/j.jhydrol.2005.07.007

Beven KJ (2012) Rainfall-runoff modelling: The primer, second edition. John Wiley and Sons Ltd, Chichester

Beven KJ, Binley AM (1992) The future of distributed models: Model calibration and uncertainty prediction. Hydrological Processes 6(3):279–298. https://doi.org/10.1002/hyp.3360060305

Beven KJ, Binley AM (2014) GLUE: 20 years on. Hydrological Processes 28(24):5897–5918. https://doi.org/10.1002/hyp.10082

Beven KJ, Freer JE (2001) Equifinality, data assimilation, and uncertainty estimation in mechanistic modelling of complex environmental systems using the GLUE methodology. Journal of Hydrology 249(1–4):11–29. https://doi.org/10.1016/S0022-1694(01)00421-8

Blöschl G, et al. (2019) Twenty-three Unsolved Problems in Hydrology (UPH) – A community perspective. Hydrological Sciences Journal 64(1):1141–1158. https://doi.org/10.1080/02626667.2019.1620507

Bock AR, Farmer WH, Hay LE (2018) Quantifying uncertainty in simulated streamflow and runoff from a continental-scale monthly water balance model. Advances in Water Resources 122:166–175. https://doi.org/10.1016/j.advwatres.2018.10.005

Bourgin F, Andréassian V, Perrin C, Oudin L (2015) Transferring global uncertainty estimates from gauged to ungauged catchments. Hydrology and Earth System Sciences 19:2535–2546. https://doi.org/10.5194/hess-19-2535-2015

Breiman L (2001) Statistical modeling: The two cultures (with comments and a rejoinder by the author). Statistical Science 16(3):199–231

Brooks SP, Gelman A (1998) General methods for monitoring convergence of iterative simulations. Journal of Computational and Graphical Statistics 7(4):434–455

Brownrigg R, Minka TP, Deckmyn A (2018) `maps`: Draw Geographical Maps. `R` package version 3.3.0. https://CRAN.R-project.org/package=maps

Clark MP, Slater AG, Rupp DE, Woods RA, Vrugt JA, Gupta HV, Wagener T, Hay LE (2008) Framework for Understanding Structural Errors (FUSE): A modular framework to diagnose differences between hydrological models. Water Resources Research 44(12):W00B02. https://doi.org/10.1029/2007WR006735

Clark MP, Nijssen B, Lundquist JD, Kavetski D, Rupp DE, Woods RA, Freer GF, Gutmann ED, Wood AW, Brekke LD, Arnold JR, Gochis DJ, Rasmussen RM (2015) A unified approach for process-based hydrologic modeling: 1. Modeling concept. Water Resources Research 51(4):2498–2514. https://doi.org/10.1002/2015WR017198

Coron L, Thirel G, Delaigue O, Perrin C, Andréassian V (2017) The suite of lumped GR hydrological models in an `R` package. Environmental Modelling and Software 94:166–171. https://doi.org/10.1016/j.envsoft.2017.05.002

Coron L, Delaigue O, Thirel G, Perrin C, Michel C (2019) `airGR`: Suite of GR Hydrological Models for Precipitation-Runoff Modelling. `R` package version 1.3.2.23. https://CRAN.R-project.org/package=airGR

Coxon G, Freer J, Westerberg IK, Wagener T, Woods R, Smith PJ (2015) A novel framework for discharge uncertainty quantification applied to 500 UK gauging stations. Water Resources Research 51(7):5531–5546. https://doi.org/10.1002/2014WR016532

Di Baldassarre G, Montanari A (2009) Uncertainty in river discharge observations: A quantitative analysis. Hydrology and Earth System Sciences 13:913–921. https://doi.org/10.5194/hess-13-913-2009

Di Baldassarre G, Laio F, Montanari A (2012) Effect of observation errors on the uncertainty of design floods. Physics and Chemistry of the Earth, Parts A/B/C 42–44:85–90. https://doi.org/10.1016/j.pce.2011.05.001

Dogulu N, López López P, Solomatine DP, Weerts AH, Shrestha DL (2015) Estimation of predictive hydrologic uncertainty using the quantile regression and UNEEC methods and their comparison on contrasting catchments. Hydrology and Earth System Sciences 19:3181–3201. https://doi.org/10.5194/hess-19-3181-2015

Dowle M, Srinivasan A (2019) `data.table`: Extension of `data.frame`. `R` package version 1.12.2. https://CRAN.R-project.org/package=data.table

Duan Q, Schaake J, Andreassian V, Franks S, Gotetie G, Gupta HV, Gusev YM, Habets F, Hall A, Hay L, Hogue T, Huang M, Leavesley G, Liang X, Nasonova ON, Noilhan J, Oudin L, Sorooshian S, Wagener T, Wood EF (2006) Model Parameter Estimation Experiment (MOPEX): An overview of science strategy and major results from the second and third workshops. Journal of Hydrology 320(1–2):3–17. https://doi.org/10.1016/j.jhydrol.2005.07.031

Edijatno, Nascimento NO, Yang X, Makhlouf Z, Michel C (1999) GR3J: A daily watershed model with three free parameters. Hydrological Sciences Journal 44(2):263–277. https://doi.org/10.1080/02626669909492221

Efstratiadis A, Koutsoyiannis D (2010) One decade of multi-objective calibration approaches in hydrological modelling: A review. Hydrological Sciences Journal 55(1):58–78. https://doi.org/10.1080/02626660903526292

Evin G, Kavetski D, Thyer M, Kuczera G (2013) Pitfalls and improvements in the joint inference of heteroscedasticity and autocorrelation in hydrological model calibration. Water Resources Research 49(7):4518–4524. https://doi.org/10.1002/wrcr.20284

Evin G, Thyer M, Kavetski D, McInerney D, Kuczera G (2014) Comparison of joint versus postprocessor approaches for hydrological uncertainty estimation accounting for error autocorrelation and heteroscedasticity. Water Resources Research 50(3):2350–2375. https://doi.org/10.1002/2013WR014185

Farmer WH, Vogel RM (2016) On the deterministic and stochastic use of hydrologic models. Water Resources Research 52(7):5619–5633. https://doi.org/10.1002/2016WR019129

Gelman A, Rubin DB (1992) Inference from iterative simulation using multiple sequences. Statistical Science 7(4):457–472

Gneiting T, Katzfuss M (2014) Probabilistic forecasting. Annual Review of Statistics and Its Application 1:125–51. https://doi.org/10.1146/annurev-statistics-062713-085831

Gneiting T, Raftery AE (2007) Strictly proper scoring rules, prediction, and estimation. Journal of the American Statistical Association 102(477):359–378. https://doi.org/10.1198/016214506000001437

Gneiting T, Balabdaoui F, Raftery AE (2007) Probabilistic forecasts, calibration and sharpness. Journal of the Royal Statistical Society: Series B (Statistical Methodology), 69(2):243–268. https://doi.org/10.1111/j.1467-9868.2007.00587.x

Haario H, Laine M, Mira A, Saksman E (2006) DRAM: Efficient adaptive MCMC. Statistics and Computing 16(4):339–354. https://doi.org/10.1007/s11222-006-9438-0

Hastie T, Tibshirani R, Friedman JH (2009) The elements of statistical learning: Data mining, inference, and prediction, second edition. Springer, New York

Hernández-López MR, Francés F (2017) Bayesian joint inference of hydrological and generalized error models with the enforcement of Total Laws. Hydrology and Earth System Sciences Discussions. https://doi.org/10.5194/hess-2017-9

Huang M, Hou Z, Leung LR, Ke Y, Liu Y, Fang Z, Sun Y (2013) Uncertainty analysis of runoff simulations and parameter identifiability in the Community Land Model: Evidence from MOPEX basins. Journal of Hydrometeorology 14(6):1754–1772. https://doi.org/10.1175/JHM-D-12-0138.1

Huard D, Mailhot A (2008) Calibration of hydrological model GR2M using Bayesian uncertainty analysis. Water Resources Research 44(2):W02424. https://doi.org/10.1029/2007WR005949

James G, Witten D, Hastie T, Tibshirani R (2013) An introduction to statistical learning. Springer-Verlag New York. https://doi.org/10.1007/978-1-4614-7138-7

Juston JM, Kauffeldt A, Montano BQ, Seibert J, Beven KJ, Westerberg IK (2012) Smiling in the rain: Seven reasons to be positive about uncertainty in hydrological modelling. Hydrological Processes 27(7):1117–1122. https://doi.org/10.1002/hyp.9625

Kavetski D, Franks SW, Kuczera G (2002) Confronting input uncertainty in environmental modelling. In: Duan Q, Gupta HV, Sorooshian S, Rousseau AN, Turcotte R (eds) Calibration of Watershed Models. AGU, pp 49–68. https://doi.org/10.1029/WS006p0049

Kavetski D, Kuczera G, Franks SW (2006a) Bayesian analysis of input uncertainty in hydrological modeling: 1. Theory. Water Resources Research 42(3):W03407. https://doi.org/10.1029/2005WR004368

Kavetski D, Kuczera G, Franks SW (2006b) Calibration of conceptual hydrological models revisited: 2. Improving optimisation and analysis. Journal of Hydrology 320(1–2):187–201. https://doi.org/10.1016/j.jhydrol.2005.07.013

Kauffeldt A, Halldin S, Rodhe A, Xu C-Y, Westerberg IK (2013) Disinformative data in large-scale hydrological modelling. Hydrology and Earth System Sciences 17:2845–2857. https://doi.org/10.5194/hess-17-2845-2013
Khatami S, Peel MC, Peterson TJ, Western AW (2019) Equifinality and flux mapping: A new approach to model evaluation and process representation under uncertainty. Water Resources Research 55. https://doi.org/10.1029/2018WR023750
Kelly KS, Krzysztofowicz R (2000) Precipitation uncertainty processor for probabilistic river stage forecasting. Water Resources Research 36:2643–2653. https://doi.org/10.1029/2000WR900061
Kim KB, Kwon HH, Han D (2018) Exploration of warm-up period in conceptual hydrological modelling. Journal of Hydrology 556:194–210. https://doi.org/10.1016/j.jhydrol.2017.11.015
Klemeš V (1986) Operational testing of hydrological simulation models. Hydrological Sciences Journal 31:13–24. https://doi.org/10.1080/02626668609491024
Koenker RW (2005) Quantile regression. Cambridge University Press, Cambridge, UK
Koenker RW (2019) `quantreg`: Quantile Regression. `R` package version 5.51. https://CRAN.R-project.org/package=quantreg
Koenker RW, Bassett Jr G (1978) Regression quantiles. Econometrica 46(1):33–50. https://doi.org/10.2307/1913643
Koenker RW, D'Orey V (1987) Computing regression quantiles. Journal of the Royal Statistical Society: Series C (Applied Statistics) 36(3):383–393. https://doi.org/10.2307/2347802
Koenker RW, D'Orey V (1994) A remark on algorithm AS 229: Computing dual regression quantiles and regression rank scores. Journal of the Royal Statistical Society: Series C (Applied Statistics) 43(2):410–414. https://doi.org/10.2307/2986030
Koenker RW, Hallock K (2001) Quantile regression: An introduction. Journal of Economic Perspectives, 15(4):43–56
Kottek M, Grieser J, Beck C, Rudolf B, Rubel F (2006) World map of the Köppen-Geiger climate classification updated. Meteorologische Zeitschrift 15(3):259–263. https://doi.org/10.1127/0941-2948/2006/0130
Koutsoyiannis D (2010) HESS Opinions: "A random walk on water". Hydrology and Earth System Sciences 14:585–601. https://doi.org/10.5194/hess-14-585-2010
Koutsoyiannis D (2011) Hurst-Kolmogorov dynamics and uncertainty. Journal of the American Water Resources Association 47(3):481–495. https://doi.org/10.1111/j.1752-1688.2011.00543.x
Koutsoyiannis D, Montanari A (2007) Statistical analysis of hydroclimatic time series: Uncertainty and insights. Water Resources Research 43(5):W05429, https://doi.org/10.1029/2006WR005592
Koutsoyiannis D, Montanari A (2015) Negligent killing of scientific concepts: The stationarity case. Hydrological Sciences Journal 60(7–8):1174–1183. https://doi.org/10.1080/02626667.2014.959959
Koutsoyiannis D, Makropoulos C, Langousis A, Baki S, Efstratiadis A, Christofides A, Karavokiros G, Mamassis N (2009) HESS Opinions: "Climate, hydrology, energy, water: recognizing uncertainty and seeking sustainability". Hydrology and Earth System Sciences 13:247–257. https://doi.org/10.5194/hess-13-247-2009
Krzysztofowicz R (1999) Bayesian theory of probabilistic forecasting via deterministic hydrologic model. Water Resources Research 35(9):2739–2750. https://doi.org/10.1029/1999WR900099

Krzysztofowicz R (2001a) The case for probabilistic forecasting in hydrology. Journal of Hydrology 249(1–4):2–9. https://doi.org/10.1016/S0022-1694(01)00420-6
Krzysztofowicz R (2001b) Integrator of uncertainties for probabilistic river stage forecasting: Precipitation-dependent model. Journal of Hydrology 249:69–85. https://doi.org/10.1016/S0022-1694(01)00413-9
Krzysztofowicz R (2002) Bayesian system for probabilistic river stage forecasting. Journal of Hydrology 268:16–40. https://doi.org/10.1016/S0022-1694(02)00106-3
Krzysztofowicz R, Herr HD (2001) Hydrologic uncertainty processor for probabilistic river stage forecasting: Precipitation-dependent model. Journal of Hydrology 249:46–68. https://doi.org/10.1016/S0022-1694(01)00412-7
Krzysztofowicz R, Kelly KS (2000) Hydrologic uncertainty processor for probabilistic river stage forecasting. Water Resources Research 36:3265–3277. https://doi.org/10.1029/2000WR900108
Kuczera G (1983) Improved parameter inference in catchment models: 1. Evaluating parameter uncertainty. Water Resources Research 19(5):1151–1162. https://doi.org/10.1029/WR019i005p01151
Kuczera G, Kavetski D, Franks S, Thyer M (2006) Towards a Bayesian total error analysis of conceptual rainfall-runoff models: Characterising model error using storm-dependent parameters. Journal of Hydrology 331(1–2):161–177. https://doi.org/10.1016/j.jhydrol.2006.05.010
Kuczera G, Renard B, Thyer M, Kavetski D (2010) There are no hydrological monsters, just models and observations with large uncertainties!. Hydrological Sciences Journal 55(6):980–991. https://doi.org/10.1080/02626667.2010.504677
Laloy E, Vrugt JA (2012) High-dimensional posterior exploration of hydrologic models using multiple-try DREAM$_{(ZS)}$ and high-performance computing. Water Resources Research 48(1):W01526. https://doi.org/10.1029/2011WR010608
Langousis A, Mamalakis A, Puliga M, Deida R (2016) Threshold detection for the generalized Pareto distribution: Review of representative methods and application to the NOAA NCDC daily rainfall database. Water Resources Research 52(4):2659–2681. https://doi.org/10.1002/2015WR018502
Lichtendahl Jr KC, Grushka-Cockayne Y, Winkler RL (2013) Is it better to average probabilities or quantiles?. Management Science 59(7):1479–1724. https://doi.org/10.1287/mnsc.1120.1667
Liu B, Nowotarski J, Hong T, Weron R (2017) Probabilistic load forecasting via quantile regression averaging on sister forecasts. IEEE Transactions on Smart Grid 8(2):730–737. https://doi.org/10.1109/TSG.2015.2437877
López López P, Verkade JS, Weerts AH, Solomatine DP (2014) Alternative configurations of quantile regression for estimating predictive uncertainty in water level forecasts for the upper Severn River: A comparison. Hydrology and Earth System Sciences 18:3411–3428. https://doi.org/10.5194/hess-18-3411-2014
Louvet S, Paturel JE, Mahé G, Rouché N, Koité M (2016) Comparison of the spatiotemporal variability of rainfall from four different interpolation methods and impact on the result of GR2M hydrological modeling—Case of Bani River in Mali, West Africa. Theoretical and Applied Climatology 123(1–2):303–319. https://doi.org/10.1007/s00704-014-1357-y
Makhlouf Z, Michel C (1994) A two-parameter monthly water balance model for French watersheds. Journal of Hydrology 162(3–4):299–318. https://doi.org/10.1016/0022-1694(94)90233-X

Mantovan P, Todini E (2006) Hydrological forecasting uncertainty assessment: Incoherence of the GLUE methodology. Journal of Hydrology 330(1–2):368–381. https://doi.org/10.1016/j.jhydrol.2006.04.046
McMillan H, Freer J, Pappenberger F, Krueger T, Clark MP (2010) Impacts of uncertain river flow data on rainfall-runoff model calibration and discharge predictions. Hydrological Processes 24(10):1270–1284. https://doi.org/10.1002/hyp.7587
McMillan H, Krueger T, Freer J (2012) Benchmarking observational uncertainties for hydrology: Rainfall, river discharge and water quality. Hydrological Processes 26(26):4078–4111. https://doi.org/10.1002/hyp.9384
Montanari A (2005) Large sample behaviors of the generalized likelihood uncertainty estimation (GLUE) in assessing the uncertainty of rainfall-runoff simulations. Water Resources Research 41(8):W08406. https://doi.org/10.1029/2004WR003826
Montanari A (2007) What do we mean by 'uncertainty'? The need for a consistent wording about uncertainty assessment in hydrology. Hydrological Processes 21(6):841–845. https://doi.org/10.1002/hyp.6623
Montanari A (2011) Uncertainty of hydrological predictions. In: Wilderer PA (eds) Treatise on Water Science 2. Elsevier, pp 459–478. https://doi.org/10.1016/B978-0-444-53199-5.00045-2
Montanari A, Brath A (2004) A stochastic approach for assessing the uncertainty of rainfall-runoff simulations. Water Resources Research 40(1):W01106. https://doi.org/10.1029/2003WR002540
Montanari A, Di Baldassarre G (2013) Data errors and hydrological modelling: The role of model structure to propagate observation uncertainty. Advances in Water Resources 51:498–504. https://doi.org/10.1016/j.advwatres.2012.09.007
Montanari A, Grossi G (2008) Estimating the uncertainty of hydrological forecasts: A statistical approach. Water Resources Research 44(12):W00B08. https://doi.org/10.1029/2008WR006897
Montanari A, Koutsoyiannis D (2012) A blueprint for process-based modeling of uncertain hydrological systems. Water Resources Research 48(9):W09555. https://doi.org/10.1029/2011WR011412
Montanari A, Koutsoyiannis D (2014) Modeling and mitigating natural hazards: Stationarity is immortal!. Water Resources Research 50(12):9748–9756. https://doi.org/10.1002/2014WR016092
Mouelhi S, Michel C, Perrin C, Andréassian V (2006a) Linking stream flow to rainfall at the annual time step: The Manabe bucket model revisited. Journal of Hydrology 328(1–2):283–296. https://doi.org/10.1016/j.jhydrol.2005.12.022
Mouelhi S, Michel C, Perrin C, Andréassian V (2006b) Stepwise development of a two-parameter monthly water balance model. Journal of Hydrology 318(1–4):200–214. https://doi.org/10.1016/j.jhydrol.2005.06.014
Nearing GS, Tian Y, Gupta HV, Clark MP, Harrison KW, Weijs SV (2016) A philosophical basis for hydrological uncertainty. Hydrological Sciences Journal 61(9):1666–1678. https://doi.org/10.1080/02626667.2016.1183009
Niel H, Paturel JE, Servat E (2003) Study of parameter stability of a lumped hydrologic model in a context of climatic variability. Journal of Hydrology 278(1–4):213–230. https://doi.org/10.1016/S0022-1694(03)00158-6
Nowotarski J, Liu B, Weron R, Hong T (2016) Improving short term load forecast accuracy via combining sister forecasts. Energy 98(1):40–49. https://doi.org/10.1016/j.energy.2015.12.142

Papacharalampous GA, Tyralis H (2018) Evaluation of random forests and Prophet for daily streamflow forecasting. Advances in Geosciences 45:201–208. https://doi.org/10.5194/adgeo-45-201-2018

Papacharalampous GA, Tyralis H, Koutsoyiannis D (2018a) One-step ahead forecasting of geophysical processes within a purely statistical framework. Geoscience Letters 5(1):12. https://doi.org/10.1186/s40562-018-0111-1

Papacharalampous GA, Tyralis H, Koutsoyiannis D (2018b) Predictability of monthly temperature and precipitation using automatic time series forecasting methods. Acta Geophysica 66(4):807–831. https://doi.org/10.1007/s11600-018-0120-7

Papacharalampous GA, Tyralis H, Koutsoyiannis D (2018c) Univariate time series forecasting of temperature and precipitation with a focus on machine learning algorithms: A multiple-case study from Greece. Water Resources Management 32(15):5207–5239. https://doi.org/10.1007/s11269-018-2155-6

Papacharalampous GA, Tyralis H, Koutsoyiannis D (2019a) Comparison of stochastic and machine learning methods for multi-step ahead forecasting of hydrological processes. Stochastic Environmental Research and Risk Assessment 33(2):481–514. https://doi.org/10.1007/s00477-018-1638-6

Papacharalampous GA, Koutsoyiannis D, Montanari A (2019b) Quantification of predictive uncertainty in hydrological modelling by harnessing the wisdom of the crowd: Methodology development and investigation using toy models. Advances in Water Resources. https://doi.org/10.1016/j.advwatres.2019.103471

Papacharalampous GA, Tyralis H, Koutsoyiannis D, Montanari A (2019c) Supplementary material for the paper "Quantification of predictive uncertainty in hydrological modelling by harnessing the wisdom of the crowd: A large-sample experiment at monthly timescale". figshare. https://doi.org/10.6084/m9.figshare.7959473.v2

Papacharalampous GA, Tyralis H, Langousis A, Jayawardena AW, Sivakumar B, Mamassis N, Montanari A, Koutsoyiannis D (2019d) Probabilistic hydrological post-processing at scale: Why and how to apply machine-learning quantile regression algorithms. Water 11(10):2126. https://doi.org/10.3390/w11102126

Papalexiou SM, Koutsoyiannis D (2013) Battle of extreme value distributions: A global survey on extreme daily rainfall. Water Resources Research 49(1):187–201. https://doi.org/10.1029/2012WR012557

Pappenberger F, Beven KJ (2006) Ignorance is bliss: Or seven reasons not to use uncertainty analysis. Water Resources Research 42(5):W05302. https://doi.org/10.1029/2005WR004820

Pappenberger F, Ramos MH, Cloke HL, Wetterhall F, Alfieri L, Bogner K, Mueller A, Salamon P (2015) How do I know if my forecasts are better? Using benchmarks in hydrological ensemble prediction. Journal of Hydrology 522:697–713. https://doi.org/10.1016/j.jhydrol.2015.01.024

Paturel JE, Servat E, Vassiliadis A (1995) Sensitivity of conceptual rainfall-runoff algorithms to errors in input data—Case of the GR2M model. Journal of Hydrology 168(1–4):111–125. https://doi.org/10.1016/0022-1694(94)02654-T

Pechlivanidis IG, Jackson BM, McIntyre NR, Wheater HS (2011) Catchment scale hydrological modelling: A review of model types, calibration approaches and uncertainty analysis methods in the context of recent developments in technology and applications. Global NEST Journal 13(3):193–214

Perrin C, Michel C, Andréassian V (2001) Does a large number of parameters enhance model performance? Comparative assessment of common catchment model structures on 429 catchments. Journal of Hydrology 242(3–4):275–301. https://doi.org/10.1016/S0022-1694(00)00393-0

Perrin C, Michel C, Andréassian V (2003) Improvement of a parsimonious model for streamflow simulation. Journal of Hydrology 279(1–4):275–289. https://doi.org/10.1016/S0022-1694(03)00225-7

Peterson RA (2017) Estimating normalization transformations with `bestNormalize`. https://github.com/petersonR/bestNormalize

Peterson RA (2019) `bestNormalize`: Normalizing Transformation Functions. `R` package version 1.4.0. https://CRAN.R-project.org/package=bestNormalize

Plummer M, Best N, Cowles K, Vines K (2006) CODA: convergence diagnosis and output analysis for MCMC. `R` news 6(1):7–11

Plummer M, Best N, Cowles K, Vines K, Sarkar D, Bates D, Almond R, Magnusson A (2019) `coda`: Output Analysis and Diagnostics for MCMC. `R` package version 0.19-3. https://CRAN.R-project.org/package=coda

Quilty J, Adamowski J, Boucher MA (2019) A stochastic data-driven ensemble forecasting framework for water resources: A case study using ensemble members derived from a database of deterministic wavelet-based models. Water Resources Research 55(1):175–202. https://doi.org/10.1029/2018WR023205

`R` Core Team (2019). `R`: A language and environment for statistical computing. `R` Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org

Ramos MH, Mathevet T, Thielen J, Pappenberger F (2010) Communicating uncertainty in hydro-meteorological forecasts: Mission impossible?. Meteorological Applications 17(2):223–235. https://doi.org/10.1002/met.202

Ramos MH, Van Andel SJ, Pappenberger F (2013) Do probabilistic forecasts lead to better decisions?. Hydrology and Earth System Sciences 17:2219–2232. https://doi.org/10.5194/hess-17-2219-2013

Ren H, Hou Z, Huang M, Bao J, Sun Y, Tesfa T, Leung LR (2016) Classification of hydrological parameter sensitivity and evaluation of parameter transferability across 431 US MOPEX basins. Journal of Hydrology 536:92–108. https://doi.org/10.1016/j.jhydrol.2016.02.042

Renard B, Kavetski D, Kuczera G, Thyer M, Franks SW (2010) Understanding predictive uncertainty in hydrologic modeling: The challenge of identifying input and structural errors. Water Resources Research 46(5):W05521. https://doi.org/10.1029/2009WR008328

Renard B, Kavetski D, Leblois E, Thyer M, Kuczera G, Franks SW (2011) Toward a reliable decomposition of predictive uncertainty in hydrological modeling: Characterizing rainfall errors using conditional simulation. Water Resources Research 47(11):W11516. https://doi.org/10.1029/2011WR010643

Romero-Cuellar J, Abbruzzo A, Adelfio G, Francés F (2019) Hydrological post-processing based on approximate Bayesian computation (ABC). Stochastic Environmental Research and Risk Assessment 33(7):1361–1373. https://doi.org/10.1007/s00477-019-01694-y

Sadegh M, Vrugt JA (2013) Bridging the gap between GLUE and formal statistical approaches: Approximate Bayesian computation. Hydrology and Earth System Sciences 17:4831–4850. https://doi.org/10.5194/hess-17-4831-2013

Sadegh M, Vrugt JA (2014) Approximate Bayesian Computation using Markov Chain Monte Carlo simulation: DREAM(ABC). Water Resources Research 50(8):6767–6787. https://doi.org/10.1002/2014WR015386
Sadegh M, Vrugt JA, Xu C, Volpi E (2015) The stationarity paradigm revisited: Hypothesis testing using diagnostics, summary metrics, and DREAM(ABC). Water Resources Research 51(11):9207–9231. https://doi.org/10.1002/2014WR016805
Sawicz K, Wagener T, Sivapalan M, Troch PA, Carrillo G (2011) Catchment classification: empirical analysis of hydrologic similarity based on catchment function in the eastern USA. Hydrology and Earth System Sciences 15:2895–2911. https://doi.org/10.5194/hess-15-2895-2011
Schaake J, Cong S, Duan Q (2006) US MOPEX data set. IAHS Publication 307:9–28
Schaake JC, Duan Q, Smith M, Koren V (2000) Criteria to select basins for hydrologic model development and testing. Preprints in: 15th Conference on Hydrology (Long Beach, California, USA, Am. Met. Soc., 10–14 January 2000), Paper P1.8
Schoups G, Vrugt JA (2010) A formal likelihood function for parameter and predictive inference of hydrologic models with correlated, heteroscedastic, and non-Gaussian errors. Water Resources Research 46(10):W10531. https://doi.org/10.1029/2009WR008933
Shmueli G (2010) To explain or to predict?. Statistical Science 25(3):289–310. https://doi.org/10.1214/10-STS330
Sikorska AE, Montanari A, D Koutsoyiannis (2015) Estimating the uncertainty of hydrological predictions through data-driven resampling techniques. Journal of Hydrologic Engineering 20(1):A4014009. https://doi.org/10.1061/(ASCE)HE.1943-5584.0000926
Solomatine DP, Shrestha DL (2009) A novel method to estimate model uncertainty using machine learning techniques. Water Resources Research 45(12):W00B11. https://doi.org/10.1029/2008WR006839
Soetaert K, Petzoldt T (2010) Inverse modelling, sensitivity and Monte Carlo analysis in R using package FME. Journal of Statistical Software 33(3):1–28. https://doi.org/10.18637/jss.v033.i03
Soetaert K, Petzoldt T (2016) FME: A Flexible Modelling Environment for Inverse Modelling, Sensitivity, Identifiability and Monte Carlo Analysis. R package version 1.3.5. https://CRAN.R-project.org/package=FME
Stedinger JR, Vogel RM, Lee SU, Batchelder R (2008) Appraisal of the generalized likelihood uncertainty estimation (GLUE) method. Water Resources Research, 44(12). https://doi.org/10.1029/2008WR006822
Széles B, Broer M, Parajka J, Hogan P, Eder A, Strauss P, Blöschl G (2018) Separation of scales in transpiration effects on low flows: A spatial analysis in the Hydrological Open Air Laboratory. Water Resources Research 54(9):6168–6188. https://doi.org/10.1029/2017WR022037
Thyer M, Renard B, Kavetski D, Kuczera G, Franks SW, Srikanthan S (2009) Critical evaluation of parameter consistency and predictive uncertainty in hydrological modeling: A case study using Bayesian total error analysis. Water Resources Research 45(12):W00B14. https://doi.org/10.1029/2008WR006825
Todini E (2004) Role and treatment of uncertainty in real-time flood forecasting. Hydrological Processes 18:2743–2746. https://doi.org/10.1002/hyp.5687
Todini E (2007) Hydrological catchment modelling: past, present and future. Hydrology and Earth System Sciences 11:468–482. https://doi.org/10.5194/hess-11-468-2007

Todini E (2008) A model conditional processor to assess predictive uncertainty in flood forecasting. International Journal of River Basin Management 6(2):123–137. https://doi.org/10.1080/15715124.2008.9635342
Tomkins KM (2014) Uncertainty in streamflow rating curves: methods, controls and consequences. Hydrological Processes 28(3):464–481. https://doi.org/10.1002/hyp.9567
Toth E, Montanari A, Brath A (1999) Real-time flood forecasting via combined use of conceptual and stochastic models. Physics and Chemistry of the Earth, Part B: Hydrology, Oceans and Atmosphere 24(7):793–798. https://doi.org/10.1016/S1464-1909(99)00082-9
Tyralis H, Koutsoyiannis D (2017) On the prediction of persistent processes using the output of deterministic models. Hydrological Sciences Journal 62(13):2083–2102. https://doi.org/10.1080/02626667.2017.1361535
Tyralis H, Papacharalampous GA (2017) Variable selection in time series forecasting using random forests. Algorithms 10(4):114. https://doi.org/10.3390/a10040114
Tyralis H, Papacharalampous GA (2018) Large-scale assessment of Prophet for multi-step ahead forecasting of monthly streamflow. Advances in Geosciences 45:147–153. https://doi.org/10.5194/adgeo-45-147-2018
Tyralis H, Dimitriadis P, Koutsoyiannis D, O'Connell PE, Tzouka K, Iliopoulou T (2018) On the long-range dependence properties of annual precipitation using a global network of instrumental measurements. Advances in Water Resources 111:301–318. https://doi.org/10.1016/j.advwatres.2017.11.010
Tyralis H, Papacharalampous GA, Burnetas A, Langousis A (2019a) Hydrological post-processing using stacked generalization of quantile regression algorithms: Large-scale application over CONUS. Journal of Hydrology 577:123957. https://doi.org/10.1016/j.jhydrol.2019.123957
Tyralis H, Papacharalampous GA, Langousis A (2019b) A brief review of random forests for water scientists and practitioners and their recent history in water resources. Water 11(5):910. https://doi.org/10.3390/w11050910
Tyralis H, Papacharalampous GA, Tantanee S (2019c) How to explain and predict the shape parameter of the generalized extreme value distribution of streamflow extremes using a big dataset. Journal of Hydrology 574:628–645. https://doi.org/10.1016/j.jhydrol.2019.04.070
Vitolo C (2017) `hddtools`: hydrological data discovery tools. The Journal of Open Source Software 2(9). https://doi.org/10.21105/joss.00056
Vitolo C (2018) `hddtools`: Hydrological Data Discovery Tools. `R` package version 0.8.2. https://CRAN.R-project.org/package=hddtools
Vogel RM (1999) Stochastic and deterministic world views. Journal of Water Resources Planning and Management 125(6):311–313
Volpi E, Schoups G, Firmani G, Vrugt JA (2017) Sworn testimony of the model evidence: Gaussian Mixture Importance (GAME) sampling. Water Resources Research 53(7):6133– 6158. https://doi.org/10.1002/2016WR020167
Vrugt JA (2016) Markov chain Monte Carlo simulation using the DREAM software package: Theory, concepts, and `MATLAB` implementation. Environmental Modelling and Software 75:273–316. https://doi.org/10.1016/j.envsoft.2015.08.013
Vrugt JA (2018) `MODELAVG`: A `MATLAB` Toolbox for postprocessing of model ensembles [preprint made available by the author]
Vrugt JA (2019) Merging models with data. Topic 6: Model averaging [presentation made available by the author]

Vrugt JA, Ter Braak CJF, Clark MP, Hyman JM, Robinson BA (2008) Treatment of input uncertainty in hydrologic modeling: Doing hydrology backward with Markov chain Monte Carlo simulation. Water Resources Research 44(12):W00B09. https://doi.org/10.1029/2007WR006720

Vrugt JA, Ter Braak CJF, Diks CGH, Robinson BA, Hyman JM, Higdon D (2009a) Accelerating Markov chain Monte Carlo simulation by differential evolution with self-adaptive randomized subspace sampling. International Journal of Nonlinear Sciences and Numerical Simulation 10(3). https://doi.org/10.1515/IJNSNS.2009.10.3.273

Vrugt JA, Ter Braak CJF, Gupta HV, Robinson BA (2009b) Equifinality of formal (DREAM) and informal (GLUE) Bayesian approaches in hydrologic modeling?. Stochastic Environmental Research and Risk Assessment 23(7):1011–1026. https://doi.org/10.1007/s00477-008-0274-y

Vrugt JA, Ter Braak CJF, Diks CGH, Schoups G (2013) Hydrologic data assimilation using particle Markov chain Monte Carlo simulation: Theory, concepts and applications. Advances in Water Resources 51:457–478. https://doi.org/10.1016/j.advwatres.2012.04.002

Wagener T, Hogue T, Schaake J, Duan Q, Gupta H, Andreassian V, Hall A, Leavesley G (2006) The Model Parameter Estimation Experiment (MOPEX): Its structure, connection to other international initiatives and future directions. IAHS Publication Series 307:339–346. https://www.osti.gov/servlets/purl/898007

Waldmann E (2018) Quantile regression: A short story on how and why. Statistical Modelling 18(3–4):1–16. https://doi.org/10.1177/1471082X18759142

Wang P, Liu B, Hong T (2016) Electric load forecasting with recency effect: A big data approach. International Journal of Forecasting 32(3):585–597. https://doi.org/10.1016/j.ijforecast.2015.09.006

Wani O, Beckers JVL, Weerts AH, Solomatine DP (2017) Residual uncertainty estimation using instance-based learning with applications to hydrologic forecasting. Hydrology and Earth System Sciences 21:4021–4036. https://doi.org/10.5194/hess-21-4021-2017

Warnes GR, Bolker B, Gorjanc G, Grothendieck G, Korosec A, Lumley T, MacQueen D, Magnusson A, Rogers J (2017) `gdata`: Various `R` Programming Tools for Data Manipulation. `R` package version 2.18.0. https://CRAN.R-project.org/package=gdata

Weijs SV, Schoups G, Van de Giesen N (2010) Why hydrological predictions should be evaluated using information theory. Hydrology and Earth System Sciences 14:2545–2558. https://doi.org/10.5194/hess-14-2545-2010

Weijs SV, van de Giesen N, Parlange MB (2013) HydroZIP: How hydrological knowledge can be used to improve compression of hydrological data. Entropy 15(4):1289–1310; https://doi.org/10.3390/e15041289

Wickham H (2007) Reshaping data with the `reshape` package. Journal of Statistical Software 21(12):1–20

Wickham H (2011) The split-apply-combine strategy for data analysis. Journal of Statistical Software 40(1):1–29

Wickham H (2016a) `ggplot2`. Springer International Publishing. https://doi.org/10.1007/978-3-319-24277-4

Wickham H (2016b) `plyr`: Tools for Splitting, Applying and Combining Data. `R` package version 1.8.4. https://CRAN.R-project.org/package=plyr

Wickham H (2018) `reshape`: Flexibly Reshape Data. `R` package version 0.8.8. https://CRAN.R-project.org/package=reshape

Wickham H, Henry L (2019) tidyr: Easily Tidy Data with 'spread()' and 'gather()' Functions. R package version 0.8.3. https://CRAN.R-project.org/package=tidyr
Wickham H, Hester J, Francois R (2018) readr: Read Rectangular Text Data. R package version 1.3.1. https://CRAN.R-project.org/package=readr
Wickham H, Chang W, Henry L, Pedersen TL, Takahashi K, Wilke C, Woo K, Yutani H (2019a) ggplot2: Create Elegant Data Visualisations Using the Grammar of Graphics. R package version 3.2.1. https://CRAN.R-project.org/package=ggplot2
Wickham H, François R, Henry L, Müller K (2019b) dplyr: A Grammar of Data Manipulation. R package version 0.8.3. https://CRAN.R-project.org/package=dplyr
Wickham H, Hester J, Chang W (2019c) devtools: Tools to Make Developing R Packages Easier. R package version 2.1.0. https://CRAN.R-project.org/package=devtools
Wilke CO (2018) ggridges: Ridgeline Plots in 'ggplot2'. R package version 0.5.1. https://CRAN.R-project.org/package=ggridges
Xie Y (2014) knitr: A comprehensive tool for reproducible research in R. In: Stodden V, Leisch F, Peng RD (eds) Implementing Reproducible Computational Research. Chapman and Hall/CRC
Xie Y (2015) Dynamic Documents with R and knitr, second edition. Chapman and Hall/CRC
Xie Y (2019) knitr: A General-Purpose Package for Dynamic Report Generation in R. R package version 1.24. https://CRAN.R-project.org/package=knitr
Xu C (2001) Statistical analysis of parameters and residuals of a conceptual water balance model–methodology and case study. Water Resources Management 15(2):75–92. https://doi.org/10.1023/A:1012559608269
Xu L, Chen N, Zhang X, Chen Z (2018) An evaluation of statistical, NMME and hybrid models for drought prediction in China. Journal of Hydrology 566:235–249. https://doi.org/10.1016/j.jhydrol.2018.09.020
Xu L, Chen N, Zhang X, Chen Z, Hu C, Wang C (2019) Improving the North American multi-model ensemble (NMME) precipitation forecasts at local areas using wavelet and machine learning. Climate Dynamics 53(1–2):601–615. https://doi.org/10.1007/s00382-018-04605-z
Ye A, Duan Q, Yuan X, Wood EF, Schaake J (2014) Hydrologic post-processing of MOPEX streamflow simulations. Journal of Hydrology 508:147–156. https://doi.org/10.1016/j.jhydrol.2013.10.055
Zeileis A, Grothendieck G (2005) zoo: S3 infrastructure for regular and irregular time series. Journal of Statistical Software 14(6):1–27. https://doi.org/10.18637/jss.v014.i06
Zeileis A, Grothendieck G, Ryan JA (2019) zoo: S3 Infrastructure for Regular and Irregular Time Series (Z's Ordered Observations). R package version 1.8-6. https://CRAN.R-project.org/package=zoo