Selecting components in a probabilistic hydrological forecasting chain: the benefits of an integrated evaluation

Joseph Bellier†**, Guillaume Bontron† and Isabella Zin†

†Université Grenoble Alpes, Grenoble INP, CNRS, IRD, IGE, Grenoble, France; ‡CNR, Lyon, France

ABSTRACT
Probabilistic streamflow forecasts are commonly issued by propagating meteorological ensemble forecasts into a hydrological model. However, a variety of components can be added to the chain to improve forecast skill. For operational centres, it may be easier to follow a modular strategy, and develop these components incrementally. However, only an integrated evaluation, based on the final streamflow forecasts, can ensure all components remain beneficial whatever the chain configuration. Also, it provides helpful guidelines about which developments to prioritise. Here, we consider two ensemble-based extensions, a meteorological grand ensemble and a hydrological multi-model, and two statistical corrections, a pre-processing and a post-processing. These components are opposed to their simpler alternative, and all configurations are evaluated on a common testbed comprising six catchments in the French upper Rhone River. Results show that all four components remain systematically beneficial, thereby validating the modular development strategy. Also, unlike other studies, neither the effect of the pre-processing nor that of the grand ensemble is found to vanish with hydrological modelling, which disputes the idea that all limitations in the input meteorological ensemble forcing can be rectified by a post-processing. However, only the post-processing can ensure the reliability of the streamflow forecasts; hence, it should be systematically prioritised.

Du choix des composants d’une chaîne de prévision hydrologique probabiliste : l’intérêt d’une approche intégrée d’évaluation

RESUME
Les prévisions de débit probabilistes sont communément obtenues par l’utilisation de prévisions météorologiques d’ensemble en entrée d’un modèle hydrologique. Cependant, à cette chaîne de modélisation sommaire, différents « composants » doivent être généralement ajoutés afin d’améliorer la qualité des prévisions. Pour un centre opérationnel, il est alors commode d’adopter une approche « modulaire », en identifiant les composants qui peuvent être développés de manière indépendante, car cela permet d’étaler dans le temps les efforts de développement. Néanmoins, seule une démarche « intégrée », portant sur l’évaluation des prévisions finales de débit, permet de s’assurer que l’ajout de chaque composant reste bénéfique quels que soient les autres maillons choisis pour former la chaîne. Par ailleurs, cela permet de hiérarchiser les développements, en opposant le coût de mise en place d’un composant, qui peut être d’ordre financier (accès aux prévisions météorologiques), humain (développement et support des algorithmes) ou informatique (ressources de calcul et de stockage), aux gains escomptés sur les performances de prévision. Dans cet article, qui conclut une série de développements méthodologiques précédemment, nous reprenons deux composants basés sur l’extension de l’approche ensemble, un modèle météorologique et un multi-modèle hydrologique, ainsi que deux corrections statistiques, le pré-traitement et le post-traitement. En activant ou désactivant ces quatre composants, nous construisons de multiples chaînes de modélisation, et évaluons les prévisions de débits obtenues avec chacune d’elles, sur six bassins versants du Haut-Rhône français. Les résultats nous permettent de valider l’approche modulaire de développement, car ils montrent que l’utilisation pour chaque maillon de la chaîne du composant le plus sophistiqué conduit bien aux meilleures performances, et ce quels que soient les autres maillons choisis. Ensuite, nous observons que les améliorations apportées au forçage météorologique, par le pré-traitement ou l’utilisation d’un grand ensemble, restent bénéfiques même lorsque combinées à un post-traitement, contestant ainsi l’idée parfois discutée dans la littérature d’un post-traitement capable seul de corriger l’ensemble des défauts du forçage météorologique. Néanmoins, le post-traitement semble être l’unique composant capable d’assurer la fiabilité des prévisions finales de débits, et par conséquent sa mise en place devrait constituer une priorité.
1. Introduction

Streamflow forecasts are highly valuable for activities such as flood warning, hydropower forecasting or water resources management. They are ordinarily produced by a modelling chain made of several components chained together, the two essentials being the meteorological forecast (most often for precipitation and temperature) and the hydrological model. Because of the inherent uncertainty in this modelling chain, many operational centres are shifting from deterministic to probabilistic forecasting, with the objective of providing streamflow forecasts with an uncertainty that correctly encompasses, in a statistical sense, the observation. This desirable property, referred to as reliability (or sometimes calibration), is crucial for decision makers to be able to fully exploit the probabilistic nature of the forecasts (Krzysztofowicz, 2001). But in addition to being reliable, we expect the probabilistic forecasts to be as sharp as possible, in order to facilitate decision-making.

Historically, the meteorological community has paved the way for probabilistic forecasting with ensemble forecasting, which consists in running a numerical weather prediction (NWP) model multiple times with slightly perturbed atmosphere initial conditions, and sometimes different modelling assumptions (Bauer et al., 2015). More generally, we refer in this paper to the ensemble approach as any method that, for a given physical process to be modelled, constructs an ensemble of simulations whose dispersion represents the uncertainty of that process. The introduction of the ensemble approach into hydrological forecasting has most often consisted in propagating meteorological ensemble forecasts, issued by a given operational centre, into a hydrological model (see Cloke & Pappenberger, 2009 for a review). But other practices exist, such as using “poor man’s ensembles”, which refers to picking multiple deterministic forecasts from independent meteorological centres (Ebert, 2001), or “grand ensembles”, which merge together ensembles issued by different meteorological centres (He et al., 2009; Pappenberger et al., 2008). The objective is to benefit from the differences between the NWP models (e.g. in resolution, physical parametrisation, assimilation or perturbation methods) to widen the range of possible meteorological scenarios. Yet a meteorological ensemble forcing only exhibits the uncertainty associated with the atmospheric processes. To encompass uncertainties linked to hydrological processes and to their modelling, a common practice consists in using multiple hydrological models instead of a single one, an approach referred to as “multi-model”. Most often, several models with different structures and modelling assumptions are selected (Thiboult et al., 2016; Vélazquez et al., 2011), although one could also consider different parameter sets for the same model (e.g. the generalized likelihood uncertainty estimation [GLUE] method, Beven & Freer, 2001). One may remark that the multi-model approach follows the same principle as the “poor man’s ensemble” (i.e. different deterministic simulations are gathered to form an ensemble), although it has never been named as such, the use of physically perturbed members in hydrological modelling being still in the early stages.

The ensemble approach is particularly attractive, as it generates a “dynamic” uncertainty that is specific to each forecast case. Moreover, ensemble forecasts are intrinsically multivariate coherent, i.e. every member presents a realistic covariability across the different catchments, lead times and eventually weather variables, since they are issued by a specific run of the NWP and hydrological models. But while operational forecasting chains very commonly include one ensemble component, the use of multiple components in series (e.g. meteorological ensemble and hydrological multi-model) is rare and has been the subject of very few studies (Thiboult et al., 2016; Zappa et al., 2011).

Despite their benefits, virtually all ensemble-based forecasting chains generate streamflow forecasts that are under-dispersive and/or biased, because of sources of uncertainty that cannot be accounted for, and sometimes systematic errors that exist in the models. Statistical corrections have thus been introduced as a more pragmatic approach for ensuring the reliability of the forecasts. We mention here the book edited by Vannitsem et al. (2019) for a recent review of the field. In the hydrological community, we refer to as pre-processing the correction of the meteorological forecasts, and post-processing the correction of the streamflow forecasts. These two components can be used together, as they play different roles in the forecasting chain. The pre-processing accounts for the meteorological uncertainty, adjusting the precipitation and temperature forecasts (and other weather variables if needed) such that the meteorological “signal” is strong enough to pass through critical thresholds in the hydrological model (e.g. fill the root zone reservoir to initiate soil water percolation). The post-processing, as the last component of the chain, accounts mainly for the hydrological uncertainty, but also eventually for all sources of uncertainties that would have been ignored by the previous components.

In both corrections, the statistical adjustment is almost systematically based on univariate methods (because of the high dimension of the forecasts), which causes the multivariate coherence that was present in the ensemble forecasts to be lost. Hence, both corrections must include a multivariate procedure that restores this coherence, using techniques such as the Schaake shuffle (Clark et al., 2004), or the ensemble copula coupling (ECC, Schefzik et al., 2013).

While the use of pre-processing and post-processing has become frequent in operational hydrological forecasting, relatively few studies (Roulin &
Vannitsem, 2015; Sharma et al., 2018; Zalachor et al., 2012) have examined how they interact with each other within a complete modelling chain, and what their respective benefits are when coupled to ensemble components of varying levels of complexity. In addition, the multivariate part of the statistical correction is sometimes loosely described, although it has been shown (Bellier et al., 2017, 2018) to strongly influence the performances of the corrections.

To reduce complexity when implementing a forecasting chain that includes both ensemble and statistical components, most developers will opt for a “modular” strategy, developing each component independently. In past studies (Bellier, 2018; Bellier et al., 2017, 2018) on the French upper Rhone River testbed, such a strategy was undertaken and skilful ensemble and statistical methods were developed at each stage of the forecasting chain: meteorological forcing, pre-processing, hydrological modelling and post-processing. However, it is still necessary to ensure that every improvement at a specific stage is beneficial, whatever the combinations of components at the other stages.

Moreover, improving the forecasts has a cost, which can be financial (access to meteorological forecasts), scientific (skills required to develop the algorithm and ensure their support) or computational (resources for running ensemble simulations). For an institution or a company implementing a forecasting chain, being able to prioritise the benefits of improving specific components thus becomes highly valuable.

In this paper, we adopt an “integrated” approach, by selecting a few relevant components at each stage of the chain and evaluating the streamflow forecasts obtained with all possible combinations of components. The objective is to validate the modular development strategy, but also to provide answers to critical questions that arise in the implementation of the forecasting chain. For instance, does the gap in skill between different meteorological forcings persist as the rest of the chain complexifies? What is the best strategy to account for the meteorological uncertainty, between a grand ensemble and a pre-processing (or both)? To account for the hydrological uncertainty, is a multi-model or a post-processing (or both) most skilful? Finally, should a pre-processing or a post-processing be implemented as a priority?

The remainder of this article is organised as follows. Section 2 presents the case study. The different components of the forecasting chain are described in Section 3, and Section 4 describes the verification metrics used to evaluate the resulting streamflow forecasts. Section 5 first presents the overall results, before focusing on specific combinations to address the questions set out above. Section 6 concludes.

2. Case study
A set of six catchments, located in the upper part of the French Rhône River, are considered: Arve (stream gauge at outlet: Bout du Monde), Valserne (Lancrans), Usses (Pont Rouge), Fier (Motz), Séran (Pont de la Thuillière) and Guiers (Belmont). Figure 1 displays their location and hypsometry, while Table 1 provides some hydrological characteristics. They are located in a mountainous area, most of them experiencing snow coverage during the winter season. Yet, their altitudinal profiles are different: the Arve basin culminates at over 4000 masl while the Usses and Séran basins do not exceed 1500 masl. Hourly streamflow records at each of the gauging stations indicated in Table 1 are available from the company CNR. The 2003–2010 period was used for calibrating and validating the hydrological models, while the 2011–2014 period was reserved for evaluating the streamflow forecasts.

For CNR, which operates a series of hydropower plants on the Rhône River, this set of catchments is an
interesting testbed, as it introduces the key problematic of the multivariate coherence of the forecasts, and in particular the spatial coherence. Indeed, due to their relative proximity but also to their differences in size and hydrological regime, their streamflow exhibits various correlation patterns, which are crucial to reproduce in the forecasts.

3. Components of the forecasting chain

In this section, we briefly describe the different forecasting chain components considered in this paper, drawing from the studies of Bellier et al. (2017, 2018) and Bellier (2018), in which these components were developed and evaluated. We therefore direct the reader to these references for the technical details. A summary of all components is presented in Table 2, along with the terminology that will be used throughout the paper. For ease of understanding, the name of the components will be bolded in the text.

By choosing one component at each stage of the chain, it is possible to form \(3 \times 2 \times 2 \times 3 = 24\) combinations. The streamflow forecasts obtained with all of these combinations will be evaluated in Section 5.

3.1. Meteorological forcing

Three components are considered for the meteorological forcing stage. The first two, named ECMWF and GEFS, are ensemble forecasts of precipitation and temperature issued by individual forecast centres. These are, respectively, the European Centre for Medium-Range Weather Forecasts (ECMWF) and the National Centers for Environmental Prediction (NCEP), whose ensemble forecasts are more commonly referred to as GEFS (Global Ensemble Forecast System). While ECMWF is often considered the most skillful global ensemble product, GEFS is widely used, partly because of its easy and free access. They contain 51 and 21 members, respectively, which are here considered exchangeable (their 50 and 20 perturbed members, respectively, are not distinguished from the control member). The archives of past operational forecasts were extracted from the TIGGE database (Park et al., 2008) over the 2007–2014 period, for the cycle 00UTC, and lead times of up to 120 h (at a 6 h time step). For the needs of lumped hydrological modelling, gridded precipitation forecasts are converted to mean areal forecasts using weighted averaging, while for temperature the value of the nearest grid point (to the catchment’s centroid) is interpolated to the catchment’s median elevation using seasonal altitudinal gradients. Note that this grid-to-catchment adaptation is not considered a form of pre-processing, but rather a necessary step given our lumped hydrological modelling system.

Finally, as a third alternative forcing we also consider a grand ensemble (grandEns) that merges ECMWF and GEFS. Note that the size of grandEns (72 members) is modest compared to grand ensembles studied by, for instance, He et al. (2009) and Pappenberger et al. (2008), who leverage the full TIGGE database and form grand ensembles of 216 members. However, the financial cost (to access the operational forecasts) and the computational cost (for downloading, storing and running all these members into the hydrological models) is likely to remain unaffordable for many hydrological forecasting centres. Thus, our 72-member grandEns is a more realistic alternative, and comparing it to ECMWF and GEFS

| Stage | Component name | Description | Number of members (output) |
|-------|----------------|-------------|-----------------------------|
| Meteorological forcing | **ECMWF** | Ensemble forecast issued by the ECMWF. | 51 |
| | **GEFS** | Ensemble forecast issued by the NCEP. | 21 |
| | **grandEns** | Grand ensemble that merges ECMWF and GEFS. | 72 |
| Pre-processing | Raw | No pre-processing. | Same as input |
| | EMOS/ECC | Univariate correction using Normal-EMOS for temperature and CSG-EMOS for precipitation. Reconstruction of the multivariate structure with ECC. | Same as input* |
| Hydrological modelling | **GRP** | Hydrological model GRP coupled with the snow-accounting module Cemaneige. | Same as input |
| | **multiMod** | Multi-model composed of GRP, TOPMODEL and ARX, all coupled to Cemaneige. | Multiplied by 3 |
| Post-processing | Raw | No post-processing. | Same as input |
| | BMA/ECC | Univariate correction using BMA. Reconstruction of the multivariate structure with ECCp-Q-TS, a variant of ECC developed specifically for streamflow forecasts. | Same as input |

*Except when coupled with grandEns, where the number of members goes back to 51 (only the ECMWF ensemble members are used for ECC).

Table 1. Characteristics of the studied basins.

| Catchment | Area (km²) | Stream gauge | MQ (m³/s)* | Q90 (m³/s)* |
|-----------|------------|--------------|------------|------------|
| Aare      | 2082       | Bour du Monde| 68         | 124        |
| Valserine | 361        | Lancesrns    | 14         | 32         |
| Usses     | 390        | Point Rouge  | 3          | 9          |
| Fier      | 1375       | Motz         | 33         | 80         |
| Séran     | 290        | Pont de la Thuillière | 6 | 14 |
| Guiers    | 609        | Belmont      | 17         | 36         |

*MQ (mean streamflow) and Q90 (90th percentile streamflow) are estimated over the 2003–2014 period.
permits us to study the problematic of whether the performances of a given meteorological ensemble can be improved – or not – by adding a potentially less skilful ensemble.

In Bellier (2018), when evaluated against observed precipitation and temperature at individual catchments and lead times, these three meteorological forcings were found to rank as follows: grandEns > ECMWF > GEFS (with A > B meaning that A performs better than B).

3.2. Pre-processing

For this stage, we compare raw (i.e., without pre-processing) to EMOS/ECC, a component that includes a univariate statistical correction coupled with a multivariate procedure for reconstructing the dependence structure. The statistical correction is based on the ensemble model output statistics (EMOS, Gneiting et al., 2005) approach, which consists in replacing, for each combination of lead time and catchment, the empirical distribution of the raw ensemble members with a parametric distribution that better captures the uncertainty of the forecast. The parameters of that distribution (e.g. location, scale) are related to statistics of the raw ensemble (e.g. mean, standard deviation) via a regression model, whose coefficients are estimated by minimising a verification metric (usually the continuous ranked probability score, CRPS) over a training data set that contains past forecasts and their corresponding observations. EMOS is a family of methods that conserve this general approach but use different parametric distributions and regression models to adapt to the specificities of various weather variables.

For temperature, we apply in this study the normal-EMOS described by Gneiting et al. (2005), which uses normal distributions along with a linear regression model (the mean and variance of the predictive normal distributions are linear functions of the ensemble mean and ensemble variance, respectively). The pre-processing is applied to the temperature forecasts over the 2011–2014 validation period, using for training the past 50 days of forecasts. The length of this “sliding” training window is a compromise between the desire to account for the temporal dependencies in the raw temperature forecast errors (which calls for a short window), and the statistical stability of the regression coefficients (which calls for a longer window).

For precipitation, which has a strongly non-Gaussian behaviour, we prefer to use the CSG-EMOS proposed by Schueuerer and Hamill (2015a), which uses censored, shifted gamma (CSG) distributions. While the gamma is commonly used for right-skewed, non-negative variables, the left-shifting with censoring at zero allows us to model precipitation amount and probability of precipitation simultaneously. The regression model is also more sophisticated than in the normal-EMOS, to account, among other things, for the heteroscedasticity in the uncertainty of the precipitation forecasts. Because of the intermittency of precipitation data, more forecasts are, however, needed for training. To increase the training period while accounting for seasonal effects, we use a monthly calibration, with the running past 4 years of forecasts (as a compromise between the size of the training period and updates of the operational forecasting systems).

The full results of the EMOS pre-processing are presented in Bellier (2018), where it is found that the meteorological forecasts after pre-processing are fairly reliable for both variables. It is important to note that the raw ensemble members have been considered exchangeable in the pre-processing. While this is a reasonable assumption when a single ensemble is being pre-processed (ECMWF or GEFS), it is sub-optimal for grandEns, whose pre-processing could probably benefit from a member weighting scheme, or a kernel dressing method such as Bayesian model averaging (BMA), which allows multi-modal distributions (see Section 3.4). One must therefore keep this limitation in mind when assessing (in Section 5) the skill of the combination grandEns-EMOS/ECC.

For reconstructing the between-catchment, lead time and weather variable (temperature and precipitation) dependence structure that has been lost through the statistical correction, we here use the ECC technique (Schefzik et al., 2013). This consists in sampling from the EMOS distributions the same number of members as in the raw forecast, and reordering them such that the rank structure is identical to that of the raw forecast members. In other words, ECC reproduces on the pre-processed ensembles the same spatial, temporal and inter-variable structure as that of the raw ensembles. Bellier et al. (2017) found the ECC technique to strongly outperform the popular Schaake shuffle method (Clark et al., 2004), which reproduces a “climatological” (and, thus, not weather-dependent) multivariate structure.

In the above-mentioned study by Bellier et al. (2017), when evaluated against observed precipitation at multiple catchments and lead times simultaneously, but also against simulated streamflow at the outlet of the water system, the two components were found to rank as follows: EMOS/ECC > raw.

3.3. Hydrological modelling

A multi-model (multiMod) is here composed of three different hydrological models: TOPMODEL (Beven, 2012), a variable contributing area model, GRP (Berthet et al., 2009), a parsimonious reservoir model, designed specifically for flood forecasting, and finally a fully statistical autoregressive exogenous
model (ARX, Remesan & Mathew, 2015) based on regression equations considering previous streamflow and precipitation as inputs, and with coefficients segmented by model state parameters. As implemented here, the three models are driven by mean areal precipitation and mean areal temperature forecasts, and assimilate the latest streamflow observations with the objective of reducing the forecast errors on the first lead times. All three models are coupled with the snow-accounting module Cemaneige (Valéry et al., 2014). Their calibration was performed in simulation mode (i.e. without streamflow assimilation), by minimising the Nash–Sutcliffe efficiency (NSE) of the square-root transformed flow. In forecast mode, the simulations are launched at 00 UTC, with an hourly time step and a maximum lead time of 120 h. Figure A1 in the Appendix shows the NSE in forecast mode obtained over the validation period.

As a simpler alternative to multiMod, we also consider the single model GRP, which has been found to outperform TOPMODEL and ARX (Figure A1). Comparing multiMod to GRP is interesting to study the benefits of merging (i.e. combining into one large ensemble) the forecast members issued by different hydrological models, compared to using only the forecast members issued by the best model. In Bellier (2018), when evaluated against observed streamflow at individual catchments and lead times, these two components were found to rank as follows: multiMod > GRP.

3.4. Post-processing

As for the pre-processing, we here compare raw (i.e. no post-processing) to BMA/ECC, a component that includes a univariate statistical correction coupled with a multivariate procedure for reconstructing the dependence structure. The univariate statistical correction is based on the BMA method, introduced by Raftery et al. (2005) and adapted by Fraley et al. (2010) to deal with multiple groups of exchangeable members. Unlike the EMOS approach that considers a single parametric distribution, the BMA replaces the empirical distribution of the ensemble forecasts with a mixture of parametric distributions, or kernels. Each kernel “dresses” a given ensemble member (bias-corrected beforehand), and is weighted according to the past skill of that member. By its construction, the BMA is particularly adapted to input ensemble forecasts that are made of groups of exchangeable members (as after the hydrological multi-model), and allows multimodal distributions. In this study, normal distributions are used as kernels, after the streamflow data (forecasts and observations) have been transformed so as to be approximately Gaussian. To leverage in the most effective way the 2011–2014 period during which the streamflow forecasts are available, we consider a cross-validation strategy where years are divided into seasons, and the forecasts over a specific year and season are post-processed using as training the forecasts over the same season of the remaining years.

To reconstruct the between-catchment and between-lead time dependence structure, we apply a variant of ECC that was proposed in Bellier et al. (2018), referred to therein as ECCp-Q-TS. In a nutshell, this variant reconstructs the same multivariate dependence structure as in the raw forecasts, but specifically accounts for the specificities of hourly streamflow forecasts. These are indeed strongly autocorrelated, but also frequently non-dispersive (i.e. all members are grouped into a single trace), because of internal thresholds in the hydrological model that need to be crossed for the meteorological “signal” to trigger a hydrologic reaction.

Bellier et al. (2018) found the ensemble streamflow forecasts output by the component BMA/ECC to be reliable, to have an adequate multivariate dependence structure and to exhibit a realistic autocorrelation. In terms of skill, when evaluated against observed streamflow at multiple catchments and lead times simultaneously, it was found that BMA/ECC > raw.

4. Verification toolbox

Probabilistic forecasts should first and foremost be reliable; that is, they should provide a forecast uncertainty that correctly encompasses, in a statistical sense, the observations. For instance, when issuing 90% prediction intervals, we should expect the observation to fall within these intervals about 90% of the time (with the finite size of the verification sample causing some sampling variability). But, ideally, these intervals should also be as small as possible, to facilitate decision-making. This is the paradigm, as first formulated by Gneiting et al. (2007), of maximising the forecast sharpness while ensuring the reliability. Endorsing this guideline, our verification toolbox must necessarily include: (i) metrics that quantify simultaneously the reliability and the sharpness (we refer to this as the “overall quality”), so that forecasting systems can be compared and ranked; and (ii) diagnostic tools capable of detecting deficiencies in the reliability, so that decision makers can determine whether or not they can fully exploit the probabilistic nature of the forecasts.

In the majority of studies that have evaluated streamflow ensemble forecasts, univariate verification tools are used to evaluate forecasts at a specific catchment and lead time. However, streamflow forecasts are likely to be used for forecasting quantities such as maxima (i.e. peak flows) at the outlet of complex water systems that involve multiple catchments, or accumulated volumes across several lead times. Consequently, we believe it is important not to restrict the study to univariate verification, but to also account for the multivariate structure. In this study, we
evaluate multivariate streamflow forecasts that include the six catchments and the lead times up to +120 h. Yet, in order to focus on the medium-term spatio-temporal dependence structure and discard from the evaluation the hourly autocorrelation structure (which is supposed to be realistic whatever the component chosen at the last stage of the forecasting chain, raw or BMA/ECC), only the 20 lead times 6, 12, 18 ... , 120 h are considered for the evaluation. The dimension of our multivariate forecasts is therefore 6 catchments × 20 lead times = 120.

Verification metrics that evaluate the overall quality of multivariate ensemble forecasts include the energy score (ES, Gneiting & Raftery, 2007) and the variogram score (VS, Scheuerer & Hamill, 2015b). Additionally, by transforming multivariate quantities to univariate ones through mathematical “functionals” that are sensitive to the dependence structure, one can also evaluate multivariate ensemble forecasts using univariate verification metrics, such as the popular CRPS. Here, a relevant functional is the sum, which can be regarded as a mathematical integration, and thus translates a series of instantaneous streamflow into a volume, a quantity of particular interest for companies such as CNR as it relates to hydropower production. Therefore, in addition to the ES and VS, we consider the score referred to as CRPS_{sum} which is the CRPS of the forecast sum of streamflow over the six catchments and the 20 selected lead times. These three scores can then be turned into skill scores (respectively denoted ESS, VSS and CRPS_{sum}) that are more easily interpretable, using a reference forecast data set. Here, we consider as reference the forecasts issued by the least sophisticated combination, namely ECMWF-raw-GRP-raw. Skill scores are positively oriented: 1 corresponds to perfect forecasts, 0 to forecasts only as good as the reference forecasts, and negative values to forecasts less skillful than the reference forecasts. The formulation and technical details relative to these scores can be found in Bellier et al. (2018).

Besides the overall quality of the forecasts we also specifically evaluate the reliability, using the rank histogram. Although it is normally used to evaluate the reliability of univariate forecasts, the rank histogram can also be used for multivariate forecasts, with the help of the same functional “sum” (over the six catchments and the 20 selected lead times). Flatness of the histogram indicates a good reliability, U-shape and ∩-shape reveal under- and over-dispersion, respectively, and upslope and downslope shapes indicate a negative and positive bias of the central tendency, respectively. Note that the potential deviations from flatness can be caused by both the univariate forecasts and the multivariate dependence structure; hence, care must be taken when drawing conclusions about which particular behaviour in the forecasts might have caused the deficiency in the reliability.

5. Results and discussion

Figure 2 presents the performances of the streamflow forecasts obtained with the 24 different forecasting chains obtained by combining the components listed in Table 2. Although the scores CRPS_{sum}, ESS and VSS do not rank

![Figure 2](image_url)
the combinations in the exact same order, it is generally found that the sophistication of the forecasting chain goes hand in hand with increasing performance, with the most sophisticated combination, grandEns–EMOS/ECC–multiMod–BMA/ECC, being the most skilful according to all three scores. Equally important, by plotting the results of Figure 2 slightly differently it is possible to show that at every stage, the more sophisticated component systematically outperforms its simpler counterpart, whatever the components chosen at the other stages. In other words, it is found that grandEns systematically outperforms ECMWF or GEFS, that EMOS/ECC systematically outperforms raw, that multiMod systematically outperforms GRP, and finally that BMA/ECC systematically outperforms raw. As reported in Section 3, the same rankings were already observed in our past studies, but the fact that they are now verified on the final streamflow forecasts, and in a systematic way, validates the modular development strategy adopted to date.

Beyond these results, it would be interesting to identify the components that are responsible for the largest gains in performance, from the perspective of maximising the cost–benefit ratio when implementing a new forecasting chain. With that in mind, in the remainder of this section we select a few specific questions and try to answer them by comparing the performances of some chosen combinations.

5.1. On the benefits of a performant meteorological ensemble forecast as input

Given the financial cost that can represent the operational access to meteorological products, hydrological forecast centres may be tempted to select less skilful but easy-to-access ensemble forecasts, and then rely on a sophisticated chain (e.g. with statistical corrections) to correct these limitations. However, prior to committing to such a strategy, one should wonder: Do the performance gaps between ensemble forecasts from different meteorological centres persist when the rest of the forecasting chain complexes?

To answer this question, we compare in Figure 3 the results obtained with ECMWF and GEFS, when used as input for forecasting chains of increasing complexity. It is found that the gap in performance between the two ensembles does not decrease much when the chain complexes. In particular, the pre-processing is not able to fully correct the limitations of the input ensemble, as shown by the skill of GEFS-EMOS/ECC that never catches up with ECMWF-EMOS/ECC, whatever components are implemented “downstream” in the chain (i.e. hydrological modelling, post-processing). Thus, when designing a forecasting chain, one should expect a systematic increase in performance when opting for more skilful meteorological ensemble forecasts.

5.2. Strategies to better account for the meteorological uncertainty

We here study two ways of better capturing the meteorological uncertainty, once a prime meteorological centre has been chosen to provide the ensemble forecasts. The pre-processing, as a first option, faces the practical constraint of requiring a long archive of sufficiently homogeneous meteorological forecasts, as atmospheric models are updated quite often and sets of reforecasts are not systematically made available to the public. The grand ensemble approach without pre-processing, as the second option, does not have such a constraint; however, it requires an operational access to other forecast centres. Thus, a question of interest is: To better account for the meteorological uncertainty, should the available ensembles be combined to form a grand ensemble, or should the best ensemble be statistically pre-processed?

As mentioned in the introduction, the idea behind pre-processing is to correct and amplify the meteorological “signal” contained in the raw forcing so that it passes through critical thresholds in the hydrological model, while the objective of the grand ensemble is to
capture other signals that would have been missed by the prime ensemble forecast. This interpretation is illustrated by the forecast example shown in Figure 4, where ECMWF-raw does not trigger any hydrological reaction. On one hand, ECMWF-EMOS/ECC manages to amplify the raw signal sufficiently such that multiple members react. On the other hand, thanks to the GEFS ensemble, grandEns-raw brings an additional signal that ECMWF had missed, and thereby suggests to the forecaster that a hydrological reaction is possible. In this particular example, taken in winter, the additional signal in grandEns-raw originates not from precipitation but rather from the temperatures, which are higher in the GEFS forecast and thus generate more liquid precipitation. Finally, combining the two approaches (grandEns-EMOS/ECC) seems beneficial here, with all members showing a hydrological reaction, as has indeed occurred.

Figure 5 compares the verification scores obtained with these different strategies, for various levels of complexity of the components downstream in the chain. For this case study we observe that, when the chain does not include a post-processing, the gain earned by the pre-processing is larger than the gain obtained with the grand ensemble. However, as soon as a post-processing is implemented, the benefit of the grand ensemble is similar to, or even larger than, the benefit of the pre-processing. It therefore seems that implementing a grand ensemble or a pre-processing is a choice that depends on the rest of the forecasting chain, while the combination of the two remains systematically beneficial.

The reliability of the streamflow forecasts obtained with these strategies is illustrated in the rank histograms of Figure 6, shown for two levels of complexity of the downstream forecasting chain. The shape of the histograms does not change much when the grand ensemble, the pre-processing, or both components are included in the chain, which indicates that these strategies have a limited effect on the reliability of the output streamflow forecasts.

5.3. Strategies to better account for the hydrological uncertainty

Propagating meteorological forecasts, even perfectly reliable ones, into a single hydrological model will systematically lead to streamflow forecasts that are under-dispersive, because of the uncertainty of hydrological modelling that has not been accounted for. We here compare two components that address this issue: the multi-model, which relies on the between-model differences (in structure and modelling assumptions) to enlarge the spectrum of possible streamflow values, and the post-processing. The former component can be costly to set up (implementing, calibrating the models) but also to run operationally (more computational resources are required because of the larger

Figure 4. Example of the 2 January 2012 forecast on the Valserine basin, for different configurations of forecasting chains regarding the grand ensemble (grandEns) and the pre-processing (EMOS/ECC) components. Lead times (hours) are on the x-axis, streamflow (m³/s) is on the left y-axis of the lower panel, and temperature (°C) and precipitation (mm/6 h) are on the left and right y-axis of the upper panel, respectively. Observations are plotted in red.
number of members), while the latter is at first glance more affordable. However, the post-processing parameters may need to be re-calibrated on a regular basis, as many components upstream in the chain can potentially be modified and thus affect the homogeneity between the forecasts to be post-processed and the archive of past forecasts. Hence, one may wonder: To account for the hydrological uncertainty, should a multi-model be implemented, or should post-processing introduced to correct the forecasts obtained with the best model?

An analogy can be made here between the comparisons of BMA/ECC vs multiMod and of EMOS/ECC vs grandEns. Indeed, the question is, again, whether it is more beneficial to correct and amplify the signal contained in the streamflow forecasts obtained from a single hydrological model, or to rely on other models to obtain different hydrological reactions. This interpretation is similarly illustrated in Figure 7, where GRP-raw happens to strongly underestimate the observed peak flow. When looking at the GRP simulated streamflow (i.e. streamflow obtained with the
observed forcings), we notice that the underestimation of the forecasts here is in large part due to the poor simulation of the hydrological model. To address this, on one hand GRP-BMA/ECC increases the dispersion of the forecast in a statistical way to account for the modelling error of GRP, but because of the relative weakness of the input signal, the dynamic of the hydrologic reaction in the output forecast is still missed. On the other hand, multiMod-raw shows members that better simulate the dynamic of the event, thanks to the additional models TOPMODEL and ARX, whose simulated streamflows show here a larger hydrologic response to precipitation than GRP. However, we observe that all three simulated streamflows here underestimate the observed peak flow, which illustrates the limitation of the multi-model strategy alone: if in some circumstances all models are biased, the multi-model ensemble forecasts will be biased too.

Figure 8 compares the verification scores obtained with the multi-model and post-processing strategies, for various levels of complexity of the upstream forecasting chain. It is found here that using the multi-model is more skillful than post-processing the forecasts issued by the best model, but only when the chain upstream includes a pre-processing. Otherwise, it is the latter strategy that is the most skillful. Finally, combining the two approaches (multiMod-BMA/ECC) leads to the best performance, whatever the complexity of the chain upstream.

The effect of multiMod and BMA/ECC on the reliability of the streamflow forecasts is illustrated in Figure 9. We observe that the multi-model alone (multiMod-raw) is not capable of producing reliable forecasts, as shown by rank histograms that exhibit an unusual shape with two spikes in the middle of the \( \cup \)-shape (four spikes in total). These spikes reflect the too-high number of cases where the dispersion of the ensemble members within each of the three hydrological models is not sufficient for the three groups of members to mix, and the observation falls in between these groups. The post-processing component, on the other hand, produces fairly reliable forecasts, as shown by the almost-flat rank histograms. These conclusions are valid whether the upstream chain is sophisticated (grandEns-EMOS/ECC, bottom histograms) or not (ECMWF-raw, top histograms). Following our guideline of first ensuring the reliability of the forecasts, we believe that the post-processing should systematically be given priority over the multi-model. This recommendation, though, applies only if the streamflow forecasts are the “final” product to be used by the forecasters. In the case that they are only an “intermediate” product of a modelling chain that includes other components further downstream (e.g. a hydrodynamic, or a reservoir system model),
another integrated evaluation would be necessary to determine which of the two strategies is most beneficial in terms of the final output product.

5.4. Pre-processing vs post-processing

In terms of overall performance, results so far have shown that it is critical for the chain to include at least one statistical correction. The last question we discuss is, therefore, Should pre-processing or post-processing be implemented as a priority?

Pre-processing reduces the systematic biases in the meteorological forecasts, and thereby the risk of “missing” hydrological events because of critical thresholds non-exceeded in the hydrological model(s). We remind the reader that these biases may concern precipitation but also temperature, the latter playing a crucial role in winter events. Pre-processing alone, though, does not account for the hydrological uncertainty, and therefore the streamflow forecasts will likely be under-dispersive. The post-processing, on the contrary, is positioned at the end of the chain and therefore guarantees the reliability of the streamflow forecasts, by eventually aggregating the remaining meteorological uncertainty with the hydrological uncertainty. However, a potential consequence of using a post-processing alone is to generate a dispersion in the streamflow forecasts that is quite
uncorrelated with the meteorological situation. The example in Figure 10 illustrates this situation, where the ECMWF-raw-GRP-BMA/ECC forecast is indeed much more dispersed than without post-processing, but this dispersion does not reflect the dynamic of the precipitation event forecasted (i.e. three distinct precipitation events), unlike the ECMWF-EMOS/ECC-GRP-raw forecast that clearly reproduces this dynamic.

Figure 11 confirms that both strategies lead to a sound improvement in the performance (compared to a chain without any statistical correction), whether or not a grand ensemble and/or a multi-model are included in the chain. However, the results for our case study show that if only one statistical correction can be implemented, the post-processing should be preferred. Another reason for choosing post-processing over pre-processing is its ability to

Figure 10. Example of the 2 January 2012 forecast on the Valserine basin, for different configurations of forecasting chains regarding the pre-processing (EMOS/ECC) and post-processing (BMA/ECC) components. The axes are described in Figure 4.

Figure 11. Comparison of the performances obtained with different configurations of forecasting chains regarding the pre-processing (EMOS/ECC) and post-processing (BMA/ECC) components.
guarantee the reliability of the streamflow forecast, as shown by the rank histograms in Figure 12 that all exhibit a relative flatness when BMA/ECC is present. On the contrary, when only EMOS/ECC is included, the histograms reveal a strong under-dispersion of the output forecasts. Finally, Figure 11 confirms that it is indeed the combination of pre-processing and post-processing that shows the best performance, a result also obtained by Sharma et al. (2018).

Before concluding, we discuss why our findings on the benefits of pre-processing are a bit different from the conclusions drawn in some previous studies. Kang et al. (2010) and Zalachori et al. (2012) indeed found the effect of pre-processing to vanish with hydrological modelling. Verkade et al. (2013) obtained similar results, and discussed several reasons for this. First, they pointed out the strong non-linearity of their study basin (the Rhine River) and thereby of their hydrological model, which is not sensitive enough to the correction and amplification of the meteorological signal by the pre-processing. It is possible that our case study basins, which are much smaller, exhibit a weaker non-linearity and therefore are more sensitive to the effect of the pre-processing. Second, similarly to Kang et al. (2010) they use, for reconstructing the multivariate dependence structure after the statistical correction, the climatology-based Schaake shuffle method instead of the weather-dependent ECC method used in this study. The use of the Schaake shuffle in the pre-processing of precipitation forecasts has been shown to perform poorly for streamflow forecasting (Bellier et al., 2018), to the point that the benefit of the whole pre-processing component disappears. Thus, the conclusions of the above-mentioned studies would perhaps have been different if a more sophisticated version of the Schaake Shuffle, or the ECC technique, was used. Finally, the quality of the raw meteorological forcings and the method used for the statistical correction play an important role in whether a pre-processing is beneficial or not, and it is possible that their specific experiment does not play out in favour of the pre-processing.

6. Conclusions

This paper follows in a line of previous publications in which different components of a probabilistic hydrological forecasting chain were independently developed and tested: two components that extend the ensemble approach, a meteorological grand ensemble and a hydrological multi-model, and two statistical corrections, a pre-processing and a post-processing. Once this "modular" development was completed, it remained to evaluate the benefits of each component (vs its simpler alternative) when integrated into complete forecasting chains of different levels of complexity. This was the objective pursued in this paper, setting up a common verification framework on the CNR testbed on the French upper Rhone River and evaluating the output streamflow forecasts from multiple combinations of components, with the use of multivariate verification tools.

First, the results have confirmed the relevance of our modular development strategy adopted to date, since every improvement (of a component) at any specific stage of the forecasting chain has been systematically found to increase the performance of the output streamflow forecasts. In other words, no combination of other stages' components was ever found to negate the gain brought by the improvement on that specific stage. This validation of the modular strategy is a crucial answer for modellers who design operational hydrological forecasting chains, as it
allows them to spread their development efforts out over time, while still expecting a performance gain after each improvement on a specific component.

The second objective of this paper was to identify which improvements are responsible for the largest gains in performance, so that modellers can set priorities in the development of the forecasting chain. The principal findings for our specific case study are the following:

(i) The choice of the meteorological ensemble forcing used as input for the forecasting chain was found to be crucial, as any gap in performance with an alternative forcing persists even if the rest of the chain is optimal.

(ii) Once a forcing is selected, both strategies of statistically correcting it (pre-processing) or supplementing it with an additional – albeit less skilful – ensemble forcing (grand ensemble) led to a sound improvement in the streamflow forecast skills. The grand ensemble seems more beneficial if the streamflow forecasts are later post-processed; otherwise, pre-processing should be given priority.

(iii) Assuming that several hydrological models are available, both strategies of statistically correcting the streamflow forecasts issued by the best model (post-processing) or merging the forecasts issued by all models (multi-model) have been found to substantially increase the skill of the streamflow forecasts. Depending on the configuration of the chain, one or the other strategy may be found most beneficial. However, only the post-processing guarantees the reliability of the output streamflow forecasts; hence, we believe it should systematically be given priority.

(iv) Regarding the choice between pre-processing and post-processing, we recommend giving post-processing the priority, again following our guideline of first ensuring the reliability of the output streamflow forecasts. Nonetheless, unlike other studies, we have found the pre-processing to remain beneficial even when combined with a hydrological multi-model and/or a post-processing, and several reasons for why this differs from the results of other studies were suggested.

Future studies with different meteorological forcings, pre-processing/post-processing methods and hydrological models, as well as over different catchments, would be beneficial to confirm these findings.

Several questions were not tackled in this paper, although we believe they deserve further investigation. For instance, how do we compose, from a pool of available meteorological forcings, the grand ensemble that leads to the most skilful streamflow forecasts? Similarly, should all available hydrological models be included in a multi-model, or only the subset that performs best? These decisions could eventually be left to the forecaster, as a way to integrate human expertise in the otherwise fully automated forecasting chain. Indeed, based on their expertise and local knowledge, skilled forecasters may be able to identify situations where some meteorological and/or hydrological models are likely to perform better than others, and therefore decide to run the chain with a specific composition of the grand ensemble and/or the multi-model, with parameters of the statistical corrections (for that specific combination) estimated ahead of time. This approach is interesting, as it allows the forecasters to become involved in the generation of the forecast, but without proceeding to manual adjustments of the individual traces, which can add great value in a deterministic framework, but here affects the homogeneity of the forecasts that is needed for the statistical corrections to perform well. The possible interaction between human expertise and automated systems in hydrological forecasting has been the subject of relatively few papers (Berthet et al., 2019; Celie et al., 2019; Pagano et al., 2016), and we believe it is an increasingly important question that deserves more study.

**Acknowledgments**

The authors thank INRAE (Antony) for sharing the GRP hydrological model, and Michael Scheuerer for the CSG-EMOS precipitation pre-processing code. They also thank the two anonymous reviewers who helped to improve the quality of this article.

**Disclosure statement**

No potential conflict of interest was reported by the authors.

**Funding**

This research was supported by grants from Labex OSUG@2020 (Investissement d’avenir – ANR 10 LABX56) and CNR.

**ORCID**

Joseph Bellier [http://orcid.org/0000-0003-1529-2512](http://orcid.org/0000-0003-1529-2512)
Isabella Zin [http://orcid.org/0000-0001-7216-6673](http://orcid.org/0000-0001-7216-6673)

**References**

Bauer, P., Thorpe, A., & Brunet, G. (2015). The quiet revolution of numerical weather prediction. Nature, 525 (7567), 47–55. [https://doi.org/10.1038/nature14956](https://doi.org/10.1038/nature14956)

Bellier, J. (2018). Prévisions hydrologiques probabilistes dans un cadre multivarié: Quels outils pour assurer fiabilité et cohérence spatio-temporelle? [PhD thesis]. Université Grenoble Alpes. In French.

Bellier, J., Bontron, G., & Zin, I. (2017). Using meteorological analogues for reordering postprocessed precipitation ensembles in hydrological forecasting. *Water Resources
Research, 53(12), 10085–10107. https://doi.org/10.1002/2017WR021245
Bellier, J., Zin, I., & Bontron, G. (2018). Generating coherent ensemble forecasts after hydrological postprocessing: Adaptations of ECC-based methods. Water Resources Research, 54(8), 5741–5762. https://doi.org/10.1029/2018WR022601
Berthet, L., Andréassian, V., Perrin, C., & Javelle, P. (2009). How crucial is it to account for the antecedent moisture conditions in flood forecasting? Comparison of event-based and continuous approaches on 178 catchments. Hydrology and Earth System Sciences, 13(6), 819–831. https://doi.org/10.5194/hess-13-819-2009
Berthet, L., Valéry, A., Garçon, R., Marty, R., Moulin, L., Puygrenier, D., Piote, O, Le Lay, M., Janet, B., & Duquesne, F. (2019). Cohérence des prévisions et place de l’expertise: Les nouveaux défis pour la prévision des crues. La Houille Blanche, 105(1), 5–12. https://doi.org/10.1051/hlb/2019001
Beven, K. J. (2012). Rainfall-runoff modelling: The primer (2nd ed.). Wiley.
Beven, K. J., & Freer, J. (2001). Equifinality, data assimilation, and uncertainty estimation in mechanistic modeling 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
Célie, S., Bontron, G., Ouf, D., & Pont, E. (2019). Apport de l’expertise dans la prévision hydro-météorologique opérationnelle. La Houille Blanche, 105(2), 55–62. https://doi.org/10.1051/hlb/2019015
Clark, M. P., Gangopadhyay, S., Hay, L., Rajagopalan, B., & Wilby, R. (2004). The Schaeke shuffle: A method for reconstructing space–time variability in forecasted precipitation and temperature fields. Journal of Hydrometeorology, 5(1), 243–262. https://doi.org/10.1175/1525-7541(2004)005<0243:TSSA-2.0.CO;2
Cloke, H. L., & Pappenberger, F. (2009). Ensemble flood forecasting: A review. Journal of Hydrology, 375(3–4), 613–626. https://doi.org/10.1016/j.jhydrol.2009.06.005
Ebert, E. E. (2001). Ability of a poor man’s ensemble to predict the probability and distribution of precipitation. Monthly Weather Review, 129(10), 2461–2480. https://doi.org/10.1175/1520-0493(2001)129<2461:AOEPME>2.0.CO;2
Frei, C., Raftery, A. E., & Gneiting, T. (2010). Calibrating multimodel forecast ensembles with exchangeable and missing members using bayesian model averaging. Monthly Weather Review, 138(1), 190–202. https://doi.org/10.1175/2009MWR3046.1
Gneiting, T., Balabdaoui, F., & Raftery, A. E. (2007). Probabilistic forecasts, calibration and sharpness. Journal of the Royal Statistical Society, 69B(2), 243–268. https://doi.org/10.1111/j.1467-9868.2007.00587.x
Gneiting, T., & Raftery, A. E. (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., Raftery, A. E., Westveld, A. H., III, & Goldman, T. (2005). Calibrated probabilistic forecasting using ensemble model output statistics and minimum CRPS estimation. Monthly Weather Review, 133(5), 1098–1118. https://doi.org/10.1175/MWR2904.1
He, Y., Wetterhall, F., Cloke, H., Pappenberger, F., Wilson, M., Freer, J., & McGregor, G. (2009). Tracking the uncertainty in flood alerts driven by grand ensemble weather predictions. Meteorological Applications, 16(1), 91–101. https://doi.org/10.1002/met.132
Kang, T.-H., Kim, Y.-O., & Hong, I.-P. (2010). Comparison of pre- and post-processors for ensemble streamflow prediction. Atmospheric Science Letters, 11(2), 153–159. https://doi.org/10.1002/asl.276
Krzyztofowicz, R. (2001). 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
Pagano, T. C., Pappenberger, F., Wood, A. W., Ramos, M. H., Persson, A., & Anderson, B. (2016). Automation and human expertise in operational river forecasting. Wiley Interdisciplinary Reviews: Water, 3(5), 692–705. https://doi.org/10.1002/wat2.1163
Pappenberger, F., Bartholmes, J., Thielen, J., Cloke, H. L., Buizza, R., & De Roo, A. (2008). New dimensions in early flood warning across the globe using grand-ensemble weather predictions. Geophysical Research Letters, 35(10), 1–7Raftery. https://doi.org/10.1029/2008GL033837
Park, Y.-Y., Buizza, R., & Leutbecher, M. (2008). TIGGE: Preliminary results on comparing and combining ensembles. Quarterly Journal of the Royal Meteorological Society, 134(637), 2029–2050. https://doi.org/10.1002/qj.334
Raftery, A. E., Gneiting, T., Balabdaoui, F., & Polakowski, M. (2005). Using Bayesian model averaging to calibrate forecast ensembles. Monthly Weather Review, 133(5), 1155–1174. https://doi.org/10.1175/MWR2906.1
Remenas, R., & Mathew, J. (2015). Hydrological data driven modelling. Springer.
Roulin, E., & Vannitsem, S. (2015). Post-processing of medium-range probabilistic hydrological forecasting: Impact of forcing, initial conditions and model errors. Hydrological Processes, 29(6), 1434–1449. https://doi.org/10.1002/hyp.10259
Scheffzik, K., Thorarinsdottir, T. L., & Gneiting, T. (2013). Uncertainty quantification in complex simulation models using ensemble copula coupling. Statistical Science, 28(4), 616–640. https://doi.org/10.1214/13-STS443
Scheuermann, M., & Hamill, T. M. (2015a). Variogram-based proper scoring rules for probabilistic forecasts of multivariate quantities. Monthly Weather Review, 143(4), 1321–1334. https://doi.org/10.1175/MWR-D-14-00269.1
Scheuermann, M., & Hamill, T. M. (2015b). Statistical post-processing of ensemble precipitation forecasts by fitting censored, shifted Gamma distributions. Monthly Weather Review, 143(11), 4578–4596. https://doi.org/10.1175/MWR-D-15-0061.1
Sharma, S., Siddique, R., Reed, S., Ahnert, P., Mendoza Zúñiga, P. A., & Mejia, A. (2018). Relative effects of statistical preprocessing and postprocessing on a regional hydrological ensemble prediction system. Hydrology and Earth System Sciences, 22(3), 1831–1849. https://doi.org/10.5194/hess-22-1831-2018
Thibault, A., Ancill, F., & Boucher, M.-A. (2016). Accounting for three sources of uncertainty in ensemble hydrological forecasting. Hydrology and Earth System Sciences, 20(5), 1809–1825. https://doi.org/10.5194/hess-20-1809-2016
Valéry, A., Andréassian, V., & Perrin, C. (2014). ‘As simple as possible but not simpler’: What is useful in a temperature-based snow-accounting routine? Part 2–Sensitivity analysis of the Cemaneige snow accounting routine on 380 catchments. Journal of Hydrology, 517, 1176–1187. https://doi.org/10.1016/j.jhydrol.2014.04.058
Vannitsem, S., Wilks, D. S., & Messner, J. (Eds.). (2019). Statistical postprocessing of ensemble forecasts. Elsevier.
Velázquez, J., Ancill, F., Ramos, M., & Perrin, C. (2011). Can a multi-model approach improve hydrological ensemble forecasting? A study on 29 French catchments using 16
hydrological model structures. *Advances in Geosciences*, 29, 33–42. https://doi.org/10.5194/adgeo-29-33-2011

Verkade, J. S., Brown, J. D., Reggiani, P., & Weerts, A. (2013). Post-processing ECMWF precipitation and temperature ensemble reforecasts for operational hydrologic forecasting at various spatial scales. *Journal of Hydrology*, 501, 73–91. https://doi.org/10.1016/j.jhydrol.2013.07.039

Zalachori, I., Ramos, M. H., Garçon, R., Mathevet, T., & Gailhard, J. (2012). Statistical processing of forecasts for hydrological ensemble prediction: A comparative study of different bias correction strategies. *Advances in Science and Research*, 8(1), 135–141. https://doi.org/10.5194/asr-8-135-2012

Zappa, M., Jaun, S., Germann, U., Walser, A., & Fundel, F. (2011). Superposition of three sources of uncertainties in operational flood forecasting chains. *Atmospheric Research*, 100(2–3), 246–262. https://doi.org/10.1016/j.atmosres.2010.12.005

**Appendix**

Figure A1 shows the calibration metric Nash–Sutcliffe efficiency of the three hydrological models in forecast mode, over the validation period.

![Figure A1. Nash–Sutcliffe efficiency (NSE) as a function of lead time for the three hydrological models in forecast mode (i.e. with streamflow assimilation), over the validation period.](image-url)