Article:

Rougé, C. orcid.org/0000-0003-1374-4992, Peñuela, A. orcid.org/0000-0001-8039-975X and Pianosi, F. orcid.org/0000-0002-1516-2163 (2023) Forecast families: a new method to systematically evaluate the benefits of improving the skill of an existing forecast. Journal of Water Resources Planning and Management, 149 (5). 04023015. ISSN 0733-9496

https://doi.org/10.1061/jwrmd5.wreng-5934

Author-produced version of the paper published in
Journal of Water Resources Planning and Management.
Original paper available at https://ascelibrary.org/journal/jwrmd5 (ASCE).
DOI:10.1061/JWRMD5/WRENG-5934

Forecast families: a new method to systematically evaluate the benefits of improving the skill of an existing forecast

Charles Rougé[1], Andrés Peñuela[2,3] and Francesca Pianosi[3]

17 November 2022

Abstract

A growing number of studies have investigated how forecast skill, i.e. predictive power, translates into forecast value, i.e., usefulness for improving forecast-informed decisions. The relationship between skill and value is widely understood to be complex and case-specific, yet few methods enable its systematic exploration using realistic forecast errors. This paper addresses this gap by proposing a single-parameter linear scaling method to generate families of synthetic forecasts with the desired skill improvements on an existing hindcast (a retrospective forecast of already-observed events). The method is applicable to any quantity for which a deterministic or ensemble hindcast is available, and generates a set of forecasts with different skill but strictly proportional errors. This like-for-like comparison preserves the auto- and cross correlations of errors, and opens the door for thorough, yet easily interpretable, explorations of the relationship between skill and value of a realistic forecast. We apply this new method to seasonal precipitation hindcasts (produced by the ECMWF-SEAS5 forecasting system) in order to explore their value for improving the management of a water supply system in the UK. The application shows that although value generally increases with skill, the skill-value relationship is not necessarily linear, and strongly depends on operational preferences and hydrological conditions (wet or dry years). It also suggests that the forecast families methodology can help water managers and forecast developers identify when a skill increase would be most valuable. This has the potential to foster productive two-way conversations between forecast producers and users.

[1]Department of Civil and Structural Engineering, The University of Sheffield, Sheffield, S1 3JD, UK. Email: c.rouge@sheffield.ac.uk

[2]Department of Agronomy (DAUCO) Unidad de Excelancia Maria de Maeztu, University of Cordoba, Spain.

[3]Department of Civil Engineering, University of Bristol, Bristol, BS8 1TR, UK.

Introduction

In recent decades, numerical weather predictions have become considerably more accurate across timescales ranging from a few hours to several months (Bauer et al., 2015; Alley et al., 2019). The water sector stands to gain considerably from these advances, as weather forecasts can be used to anticipate changes in water demand and availability, and thus support operational decisions for flood and drought mitigation, and for effective allocation of water fluxes to agriculture, domestic consumption, hydropower production, etc. The actual uptake of weather forecasts by water managers varies across countries, sectors (e.g. water supply or hydropower production) and with lead time. For example, and even though examples exist (e.g., Fleming et al., 2021), forecasts with a lead time of over a month are still rarely used by water managers in the United States (Whateley et al., 2015; Turner et al., 2020) and in Europe (Bruno Soares and Dessai, 2016; Bruno Soares et al., 2018). This is often traced back to the uncertainty and inaccuracy of these forecast products outside the tropics (Jackson-Blake et al., 2022), but also to the challenges for managers to trial forecasts in low-stakes situations (Whateley et al., 2015).

The idea that forecast inaccuracy is a key barrier to uptake by water managers suggests a straightforward, or even proportional, relationship between a forecast's ability to inform water management decisions – its *value* – and its predictive ability or *skill*. Similar assumptions of a proportional relationship underpin efforts to increase forecast skill as a way to improve their value to end-users (e.g., Crochemore et al., 2020). On a graph relating value gains to skill gains (Figure 1.a) from a benchmark forecast (0 on the x-axis) to a perfect forecast (1), this proportional relationship corresponds to the gray line. Note that skill is a property of the forecast alone, whereas value depends not only on the quality of a forecast, but also on the physical configuration and on the operational objectives of the water resource system it is applied to (Figure 1.b). The presence of multiple controlling factors means that the skill-value relationship is not necessarily linear (Turner et al., 2017; Doering et al., 2021). For instance, if a forecast's value increased faster than its skill (similar to the blue line in Figure 1.a), then even slight improvements in forecast skill might make them much more useful for water managers.

In practical terms, the value associated with forecast information is computed by comparing an operational strategy that accounts for this information to another that does not. Early works following this approach include for example Pereira et al. (1984). Applications have increased recently to exploit improving hydro-meteorological forecasting systems (e.g., Ficchì et al., 2016; Anghileri et al., 2016; Nayak et al., 2018; Anghileri et al., 2019; Peñuela et al., 2020; Yang et al., 2020; Giuliani et al., 2020). A limitation of these studies is that they provide an assessment of a specific forecast product applied for a specific water system, thereby providing one estimate of the value for one forecast product of given skill. Aside from comparing the value gain associated with the forecast to that of a perfect forecast, these approaches offer limited opportunities of exploring the horizontal range in Figure 1.a.

Forecast generation has been introduced in the water resource literature precisely to allow a more comprehensive exploration of the link between skill and value. For a given management problem, the idea is generally to assess value of synthetic forecasts generated at different levels of skill. However, attempts at forecast generation to date have relied on idealized distributions of the error between forecast and actual variables. Early attempts added an error term to hydroclimatic variables (e.g., Maurer and Lettenmaier, 2004; Sankarasubramanian et al., 2009), an approach theoretically debunked recently

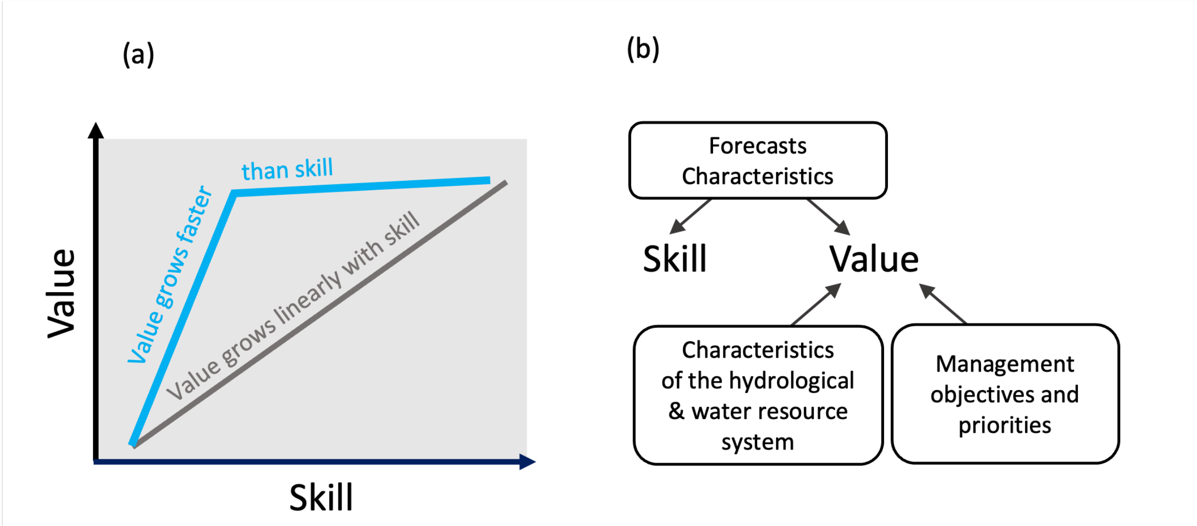

Figure 1: Left: conceptualisation of the skill-value relationship. Right: factors affecting skill and value of forecasts.

for ignoring the statistical dependence structure of the forecast generation problem Lamontagne and Stedinger (2018). On the other hand, the alternative proposed by Lamontagne and Stedinger (2018) only supports the generation of a single deterministic forecast per time series. Neither of these approaches accounts for the fact that forecasts will be updated at shorter lead times (e.g., 1 May and 1 June forecasts of 1 September streamflow will often be different). A martingale model of forecast evolution has been proposed for water resource applications to overcome this challenge, and provides a systematic understanding of the relationship between skill and value (Zhao et al., 2011, 2013). For instance, this has yielded robust insights on the superiority of ensemble forecasts over deterministic ones (Zhao and Zhao, 2014), but also warnings on the complex relationship between skill and value (Turner et al., 2017).

Despite their growing sophistication, all these efforts make the assumption that forecast errors follow idealized probability distributions. This makes these generation techniques difficult to relate to the performance of actual forecasts. Recently, Cassagnole et al. (2021) have sidestepped this problem by using perturbed streamflows, e.g., by adding bias or dispersion to obtain synthetic forecasts that lead to easily interpretable results on the relationship between forecast quality and value. While relating to actual forecasts only indirectly, this work shows the value of simple but well-conceived numerical experiments in exploring the skill-value relationship. Another noteworthy effort is a direct statistical generation approach using the empirical error correlation structure of multivariate forecasts with multiple lead times (Brodeur and Steinschneider, 2021). This approach generates synthetic forecasts that are statistically difficult to distinguish from an existing hindcast, but does not enable in itself to explore the relationship between skill and value.

More broadly, these generation methods apply time-averaged measures of forecast quality uniformly. In reality, these measures hide variations in skill depending on factors such as seasonality or flow regimes

(Pappenberger et al., 2015), themselves reflecting how well the production of a forecast captures hydro-climatic processes that play a role at different times. This is why this paper introduces a generation method meant to preserve the temporal variations in forecast quality that can be observed in actual forecasts.

This work contributes a single-parameter linear scaling approach to systematically explore the link between the skill of an existing forecast and its value to operational water resource management. The approach essentially creates a family of skill-differentiated synthetic forecasts while strictly preserving the error correlation structure of the original forecasts. It is applicable to any situation where a hindcast (forecast data on events for which the actual outcome is known) is available. We apply it to seasonal weather forecasts, in particular ECMWF SEAS5 precipitation hindcasts (Stockdale et al., 2018; Johnson et al., 2019), and explore their value for improving the operation of the UK supply water system described in Peñuela et al. (2020).

Generating a forecast family: methodology

This section sets the context and formalises the problem of constructing a synthetic forecast of desired skill for a subset of a hydro-climatic dataset, with realistic forecast errors. Specifically, we want to preserve the error correlation between the synthetic forecast and all other hindcats available within the dataset. We then propose our new method to solve this problem and its implementation for several common measures of skill for both deterministic and ensemble forecasts. Finally, we explain how the approach can be generalised into the generation of synthetic forecasts families, where each family member regroups forecasts of multiple variables issued at multiple dates, and with the desired skill.

Context and general definitions

Let us consider an observational single- or multi-location (e.g., gridded) time series dataset containing any combination of hydro-climatic variables (precipitation, temperature, streamflow, etc). Let us also consider the retrospective forecasts (or hindcasts, though we will use the term forecast in the remainder of this section) that exist for that dataset. These forecasts are issued periodically in time and for any subset of the dataset, and make predictions within a given time window in the future.

Let us now focus on one forecast $F = (F_t)_{1 \leq t \leq T}$ of a time-series of a single variable at a single location, $X = (X_t)_{1 \leq t \leq T}$ – so that each X_t is a scalar. We will discuss the generalisation to multiple variables in the last part of this Methodology section. F can be either deterministic – a single prediction for each point X_t – or an ensemble – N predictions for each point. Forecasts are commonly characterised by the error ε with respect to the quantity they are trying to predict, defined $\forall t$ as: (Lamontagne and Stedinger, 2018; Brodeur and Steinschneider, 2021):

$$X_t = F_t + \varepsilon_t \tag{1}$$

Where a benchmark forecast F^0 exists, the quality of the forecast F can be evaluated using an aggregated measure of how the errors of F compare with those of F^0, noted ε_t^0 thereafter. This measure is called the skill score and formally defined by:

$$S_f(F; X, F^0) = 1 - \frac{f(F; X)}{f(F^0; X)} \tag{2}$$

where $f(.; X)$ measures the difference between a forecast and the time series X it tries to predict. Therefore $f(X; X) = 0$, and f grows as the predictive ability decreases. As long as the forecast F is imperfect, i.e., it contains a level of error, the term $f(F; X)$ is greater than 0 and the skill $S_f(F; X, F^0)$ is smaller than 1. A skill value between 0 and 1 means that the forecast F has better performance than the benchmark F^0 as a predictor of X, whereas a negative skill means that F^0 is better than F.

Problem formulation

The aim of our methodology is to generate forecasts F of X with desired skill $S_f(F; X, F^0) = S$, and do that in a way that preserves the correlation structure of the benchmark's forecast error. This means that the errors ε_t and ε_t^0 should not only have the same auto-correlation structure, but also the same cross-correlation coefficients with the errors of other forecasts across our entire hydro-climatic dataset. Given that linear transformations of a time series preserve its auto- and cross-correlations coefficients, the simple solution we propose for this is to relate the errors from F^0 and F with a single linear scaling factor $k > 0$, so that $\forall t$:

$$\varepsilon_t = k\, \varepsilon_t^0 \tag{3}$$

By definition in equation (1) of the error terms ε_t^0 and ε_t, equation (3) is equivalent to:

$$F_t = (1 - k)\, X_t + k\, F_t^0 \tag{4}$$

and the latter equation enables one to easily generate a synthetic forecast if we have an equation linking the desired skill score S with the scaling factor k. We will now derive such relationships for three commonly used definitions of skill: two apply to deterministic forecasts and one to ensemble forecasts.

Deterministic forecasts

Two of the most common skill scores for a deterministic forecast F are based on the following performance measure f_α with $\alpha \in \{1, 2\}$:

$$f_\alpha(F; X) = \frac{1}{T} \sum_{i=1}^{T} |\varepsilon_t|^\alpha \tag{5}$$

where $\alpha = 1$ defines the mean absolute error (MAE), whereas $\alpha = 2$ defines the mean squared error (MSE). Using the definition of the scaling factor $k > 0$ from equation (3) enables us to directly relate the performance of F to that of a benchmark F^0:

$$f_\alpha(F; X) = k^\alpha\, f_\alpha(B; X) \tag{6}$$

The definition of forecast skill in equation (2) then leads to:

$$S_f(F; X, F^0) = 1 - k^\alpha \tag{7}$$

To generate a synthetic forecast of skill S while preserving the benchmark's error correlation structure, one can then apply equation (4) with the linear scaling factor k computed by:

$$k = (1 - S)^{1/\alpha} \tag{8}$$

Ensemble forecasts

A common performance measure for ensemble forecasts is the continuous ranked probability score (CRPS) (Pappenberger et al., 2015; Slater and Villarini, 2018; Peñuela et al., 2020). For the probabilistic forecast $F_t = (F_{t,i})_{1\leq i\leq N}$ of a single point X_t, it is defined as follows (Hersbach, 2000):

$$CRPS(F_t; X_t) = \int_{-\infty}^{\infty} \left(\left[\frac{1}{N} \sum_{i=1}^{N} H(z - F_{t,i}) \right] - H(z - X_t) \right)^2 dz \tag{9}$$

where H is the Heaviside function:

$$H(z - a) = \mathbb{1}(z \geq a) = \begin{cases} 0 & \text{if} \quad z < a \\ 1 & \text{if} \quad z \geq a \end{cases} \tag{10}$$

Then, the overall forecast performance is defined as the average CRPS over time:

$$CRPS(F; X) = \frac{1}{T} \sum_{t=1}^{T} CRPS(F_t; X_t) \tag{11}$$

The continuous ranked probability skill score (CRPSS) based on CRPS relates F to a benchmark F^0:

$$CRPSS(F; X, F^0) = 1 - \frac{CRPS(F; X)}{CRPS(F^0; X)} \tag{12}$$

Based on what we just did for deterministic forecasts, our goal is to relate the CRPS of F to that of F^0 using scaling factor k. Equations (6) to (8) show that then, it will be straightforward to relate k with a desired skill level S. The analytical work on the CRPS by Hersbach (2000) shows that if values within an ensemble are ordered (either in increasing or decreasing order) then the CPRS definition from equation (9) becomes:

$$CRPS(F_t; X_t) = \sum_{i=0}^{N} \left(\int_{F_{t,i}}^{F_{t,i+1}} \left[\frac{i}{N} - H(z - X_t) \right]^2 dz \right) \tag{13}$$

adopting the conventions $F_{t,0} = -\infty$ and $F_{t,N+1} = \infty$. We can decompose $CRPS(F_t; X_t)$ into a term with values of z smaller than X_t (so that $H(z - X_t) = 0$), and values of z larger than X_t (so that

$H(z - X_t) = 1$):

$$CRPS(F_t; X_t) = CRPS^-(F_t; X_t) + CRPS^+(F_t; X_t) \tag{14}$$

where we set (keep in mind that $\int_a^b f \equiv 0$ when $a \geq b$):

$$\begin{cases} CRPS^-(F_t; X_t) = \sum_{i=0}^{N} \left(\int_{F_{t,i}}^{X_t} \left[\frac{i}{N}\right]^2 dz - \int_{F_{t,i+1}}^{X_t} \left[\frac{i}{N}\right]^2 dz \right) \\ CRPS^+(F_t; X_t) = \sum_{i=0}^{N} \left(\int_{X_t}^{F_{t,i+1}} \left[\frac{i}{N} - 1\right]^2 dz - \int_{X_t}^{F_{t,i}} \left[\frac{i}{N} - 1\right]^2 dz \right) \end{cases} \tag{15}$$

Note that all integral terms above are of the form $\pm \int_{X_t}^{F_t} C$ where C is a constant. Therefore, they are equal to $\pm C(X_t - F_t) = \pm C\varepsilon_t$ according to the definition of the error in equation (1). Then, noting m the ensemble member such that $F_{t,m} \leq X_t < F_{t,m+1}$, and $\varepsilon_{t,i}$ the errors for each ordered ensemble member, the above equations (15) are equivalent to:

$$\begin{cases} CRPS^-(F_t; X_t) = \sum_{i=0}^{m} \left(\frac{i}{N}\right)^2 (\varepsilon_{t,i} - \varepsilon_{t,i+1}) \\ CRPS^+(F_t; X_t) = \sum_{i=m+1}^{N} \left(\frac{N-i}{N}\right)^2 (\varepsilon_{t,i} - \varepsilon_{t,i+1}) \end{cases} \tag{16}$$

Replacing $\varepsilon_{t,i}$ with $k\,\varepsilon^0_{t,i}$ then leads, thanks to the linearity of the sum operator, to the desired relationship between performance of F and F_0:

$$CRPS(F; X) = k\,CRPS(F^0; X) \tag{17}$$

The definition of skill based on CPRSS then becomes:

$$CRPSS(F; X, F^0) = 1 - k \tag{18}$$

This means that it is possible to directly define a single coefficient k to obtain the desired skill S of an ensemble forecast:

$$k = 1 - S \tag{19}$$

We get the same result as when defining skill from MAE in the deterministic case. This is logical given that the CRPS is equal to the MAE in the case where all ensemble members have the same value (Hersbach, 2000). Therefore, it generalises the MAE measure from deterministic to ensemble forecasts. To obtain an ensemble forecast of desired skill S from an existing benchmark F^0 with N members, and preserve the error correlation structure, it is enough to apply equation (4) to each of the N members of the original (benchmark) ensemble to obtain a synthetic ensemble of same size:

$$\forall i \in \{1, 2, \dots N\},\ F_{t,i} = (1 - k)\,X_t + k\,F^0_{t,i} \tag{20}$$

Constructing forecast families

In this section, we have demonstrated how to create a forecast of a desired skill S using a linear scaling factor $k > 0$ as a combination of a time series of observations and an available forecast. The new forecast has the same error correlation structure as the original: same error auto-correlation but also same correlation coefficients with the errors of forecasts of other hydro-climatic variables. A direct consequence of this is that the linear scaling can be applied to an arbitrary set of forecast error time series at once without affecting error correlation structure: spatially distributed forecasts, forecasts issued regularly through time, forecasts of different variables. The only constraint is that the same measure of skill needs to be used for all rescaled forecasts. Then, by applying different linear scaling factors $(k_1, k_2, \ldots, k_M)$, one can construct a forecast family of M members, where each member is a synthetic forecast generated from the chosen set of forecasts and the observations they predict.

Each choice of linear scaling factor k also rescales error bias and standard deviation; in the case of ensemble forecasts where skill is computed with CRPSS, increasing skill also means decreasing the variance in the N ensemble member forecast values.

Case-study application

To demonstrate our synthetic forecast generation approach and its value to the exploration of the skill-value relationship, we use a real-world decision-making problem of a pumped-storage water supply system in the UK. A detailed description of the system and associated simulation model can be found in Peñuela et al. (2020). The system (Figure 2.a) features two reservoirs that are used to support residential water demand, with a filling period from November to March. Ideally, by 1 April the reservoirs should be at maximum storage to ensure meeting the summer demand. A pumping station can be used to enhance the filling of Reservoir 1. Water abstractions from that reservoir can then be sent to the demand node, though this also incurs pumping costs. In contrast, meeting demand using Reservoir 2 water does not consume energy, with the caveat that abstracting too much during the filling period will reduce the amount of water available during the summer.

Seasonal forecasts could be used to anticipate hydrological conditions throughout the filling period (November to March included), and thus decide when pumping is needed. This work builds on the forecast-informed optimization workflow used by Peñuela et al. (2020), where daily ensemble rainfall and temperature forecasts provided by the ECMWF seasonal forecasting system SEAS5 (Stockdale et al., 2018; Johnson et al., 2019) for the grid cell (36 km) covering the area of study were used to force the lumped HBV rainfall-runoff model (Bergström, 1995) and generate inflow forecasts to the two reservoirs. A reservoir simulation-optimisation model was then used to reconstruct how reservoir operators could have optimally used the ensemble inflow forecasts during the reservoirs' filling period of winters 2005-2006 through 2015-2016. Comparison of the performance of forecast-informed decisions with that of rule-based operations – both against the flows that actually occurred – leads to quantifying forecast value.

In this study, we add one more step, by generating a set of synthetic rainfall and temperature forecast families from the bias-corrected ECMWF-SEAS5 hindcast, used as a benchmark. This results in the workflow presented in Figure 2.b. Skill in the forecast family goes from 0 (forecast identical to the

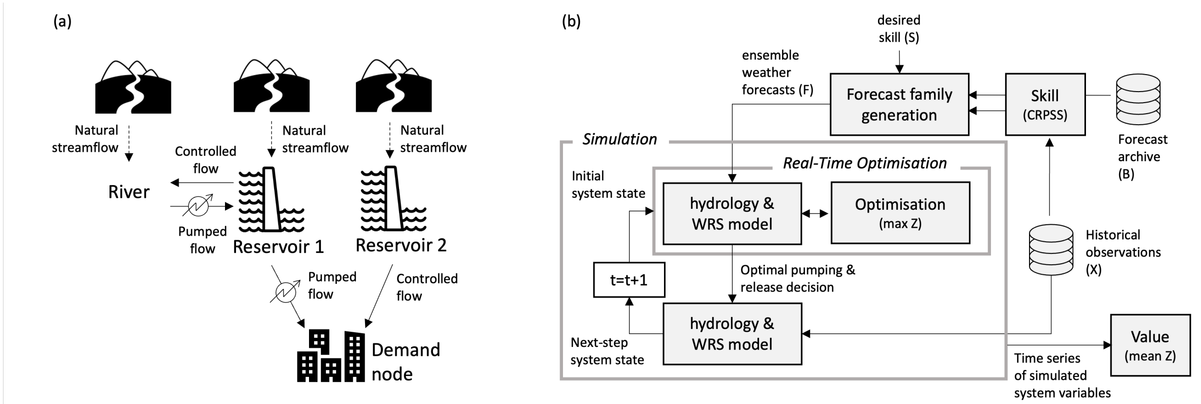

Figure 2: Left: schematic of the hydrological and water resource system used as a case study. Right: schematic of the modelling approach used to quantify the value of synthetic forecast families.

bias-corrected ECMWF-SEAS5 hindcast) to 1 (perfect forecast) by increments of 0.2, with the same skill for rainfall and temperature. The hindcasts and hence, the synthetic forecast families are daily data, with a lead time of 7 months, an ensemble size of 25 members, and a monthly issue frequency. For each generated synthetic forecast family, we generate the corresponding daily inflow forecasts and then run our simulation-optimisation model over the period 2005-2016 to quantify forecast value. The model updates its decisions every week based both on new observations and on the seasonal forecast update if one is available.

Similarly to Peñuela et al. (2020), we consider two objectives: 1) to maximize *resource availability* (RA), quantified at the end of the filling period by the storage in the two reservoirs, and 2) to minimise total *pumping costs* (PC) over the filling period. Meeting the water demand throughout the filling period is included as a constraint to the optimization. Resource Availability is expressed in percentage of total capacity, while Pumping Cost is expressed in GBP. The two objectives are aggregated into the single objective (Z), over the period 2005-2016 (the period for which ECMWF hindcasts were available):

$$Z = RA - \frac{1}{w}PC \tag{21}$$

where the weight w (GBP) modulates the relative importance given to the two objectives. Maximization of Z is performed by the Pymoo single-objective Genetic Algorithm (Blank and Deb, 2020). Similar to Peñuela et al. (2020), in this work we also analyse five possible operational preferences. The corresponding values of the weight w are given in Table 1. In the *bal* case, the value of w leads to the most balanced solution between the two objectives. This weight is multiplied or divided by two to represent the cases of prioritising either resource availability (*rap*) or pumping savings (*psp*). In the last two cases (*rao* and *pso*), the weight from *bal* is multiplied or divided by ten to establish a hierarchy where the secondary objective is only improved when there is little room left to improve the primary.

Operational preferences	Acronym	Weight w (GBP)
Resource availability only	*rao*	225,000
Resource availability prioritized	*rap*	45,000
Balanced objectives	*bal*	22,500
Pumping savings prioritized	*psp*	11,250
Pumping savings only	*pso*	2,250

Table 1: Weight given to the two objectives in the five cases of operational preferences explored in this case-study.

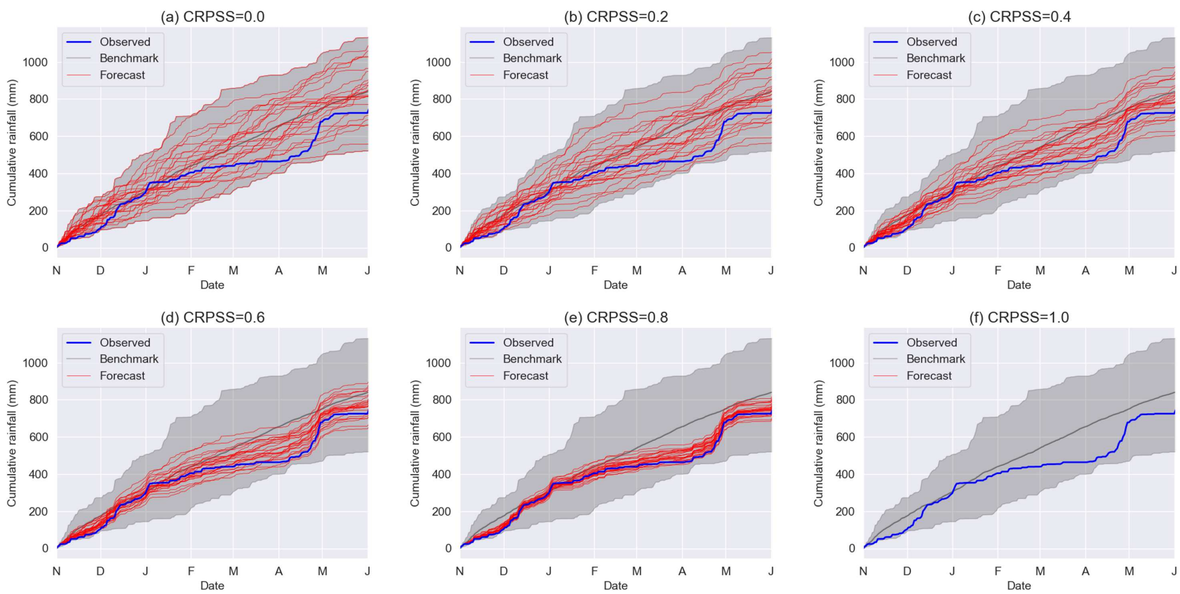

Figure 3: Forecast ensembles as the skill evolves from the original bias-corrected ECMWF forecast (CRPSS=0 as computed with this forecast as benchmark) to a perfect forecast. This is a 7-month seasonal forecast for rainfall issued on 1 November 2011.

Results

In this section, we first illustrate the generation of forecast families with an example from the case-study application. After that, we present results on the relationship between skill and value, and the insights this provides in terms of how forecast improvements can impact system operations and create value for stakeholders.

Generation of forecast families

Figure 3 visually illustrates forecast families by using the example of the bias-corrected ECMWF ensemble rainfall forecast issued for the study area on 1 November 2011 as a benchmark. Starting from the original ensemble (panel (a)), the skill defined using the benchmark goes from 0 (skill of the benchmark against itself) to 1 (skill of a perfect forecast) by increments of 0.2. In each panel, each ensemble member is a linear combination of the benchmark and the perfect forecast with respective weights $1 - CRPSS$ and $CRPSS$. The weight $1 - CRPSS$ is also the value of the parameter k which rescales the error of each ensemble member as a fraction of the original error (see equation 19). This provides a visual understanding of the evolution of an ensemble forecast as skill increases in a way that reduces the magnitude of errors without affecting their timing. Improved ensemble skill as defined by CRPSS requires both increased accuracy, as each ensemble member is closer to the actual value, and reduced ensemble variance. Note for instance how with CPRSS=0.8 (panel (e)), the average of the benchmark ensemble (in grey) is distinct from the rescaled ensemble (in red) at lead times of 5-6 months where the observed rainfall (blue) turned out to be much lower than anticipated. Illustration of how the method would apply to deterministic forecasts is presented in Appendix A.

Relationship between skill, value and operations

The relationship between skill and value for each of the five operational preferences defined in Table 1 is shown in panel (a) of Figure 4. The Figure plots value as a percentage of the perfect forecast value, i.e., as a percentage of the value of perfect climate information (corresponding to CRPSS=1 and k=0). Recall that forecast value is typically defined as the performance gain from no-forecast operations, i.e., increase of resource availability and pumping cost savings over the study period (the 11 winters from 2005-06 to 2015-16). This scaling enables us to compare value gains as skill increases with different operational priorities. Positive (resp. negative) values mean enhancement (resp. deterioration) of the performance with respect to the no-forecast operations. Figure 4.a clearly shows that value increases with the skill of the ensemble forecast, but this increase is not linear and strongly varies with the operational preferences. For instance, the initial 0.2 skill gain leads to almost no value gain with balanced preferences (*bal*), whereas it leads to significant gains with the *rao*, *rap* and *psp* preferences. The *pso* operational preference represents an extreme case where value is essentially insensitive to skill, as prioritising energy cost savings leads to pumping rarely being used, no matter what the forecast (and its skill) is. Also note that simply using the existing ECMWF-SEAS5 forecast provides at least 84% (with the *rap* preference) of the value of perfect information, with skill improvements representing minor gains in comparison.

To further analyse the skill-value relationship, the other two panels in Figure 4 show the value gains (still as a percentage of perfect forecast value gains) for the two individual objectives of maximising resource availability (panel (b)) and maximising pumping cost savings (panel (c)). Looking at the *bal* operational preferences (black line), we see that the overall gains in aggregate value as skill increases (seen in the top panel (a)) have been obtained by large increases in resource availability (panel (b)) at the expense of slight decreases in pumping savings (panel (c)). This demonstrates the complex and largely unanticipated effects that forecast skill improvements can have in multipurpose systems. Figure 4.b and c also show that when there is a strong preference for one objective (the *rao* and *pso* cases), the main gains in value with skill improvements are obtained on the secondary objective (pumping costs for *rao* and resource availability for *pso*) rather than the primary one. This is because optimised decisions focus on maximising the primary objective at the expense of the secondary one. This is why the primary objective is relatively insensitive to forecast skill whereas the secondary objective is were gains in skill are more effectively translated into performance gains.

We can further explore the relationship between gains in forecast skill and value by analyzing

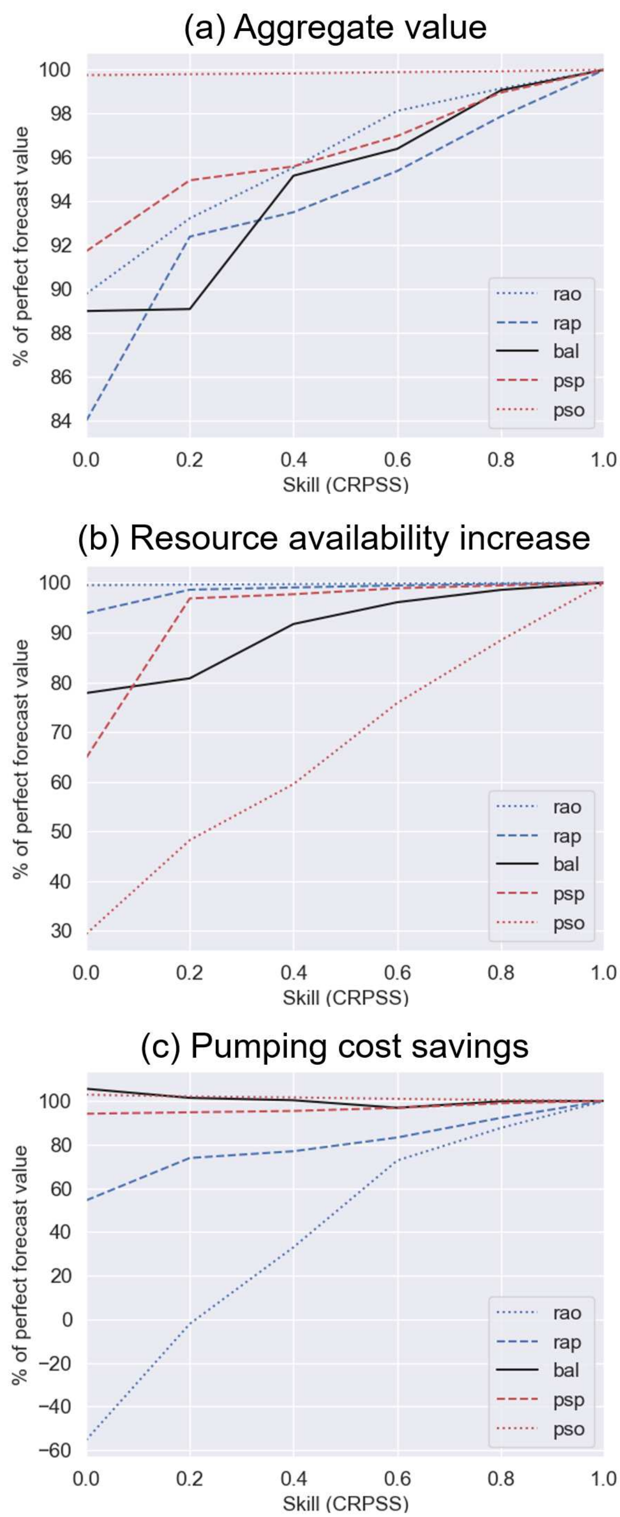

Figure 4: Skill-value relationship for five operational preferences. (a) aggregate value (expressed as percentage of the perfect forecast value); (b) increase of resource availability compared with no-forecast (percentage of perfect forecast value); and (c) pumping cost savings compared with no-forecast (percentage of perfect forecast value).

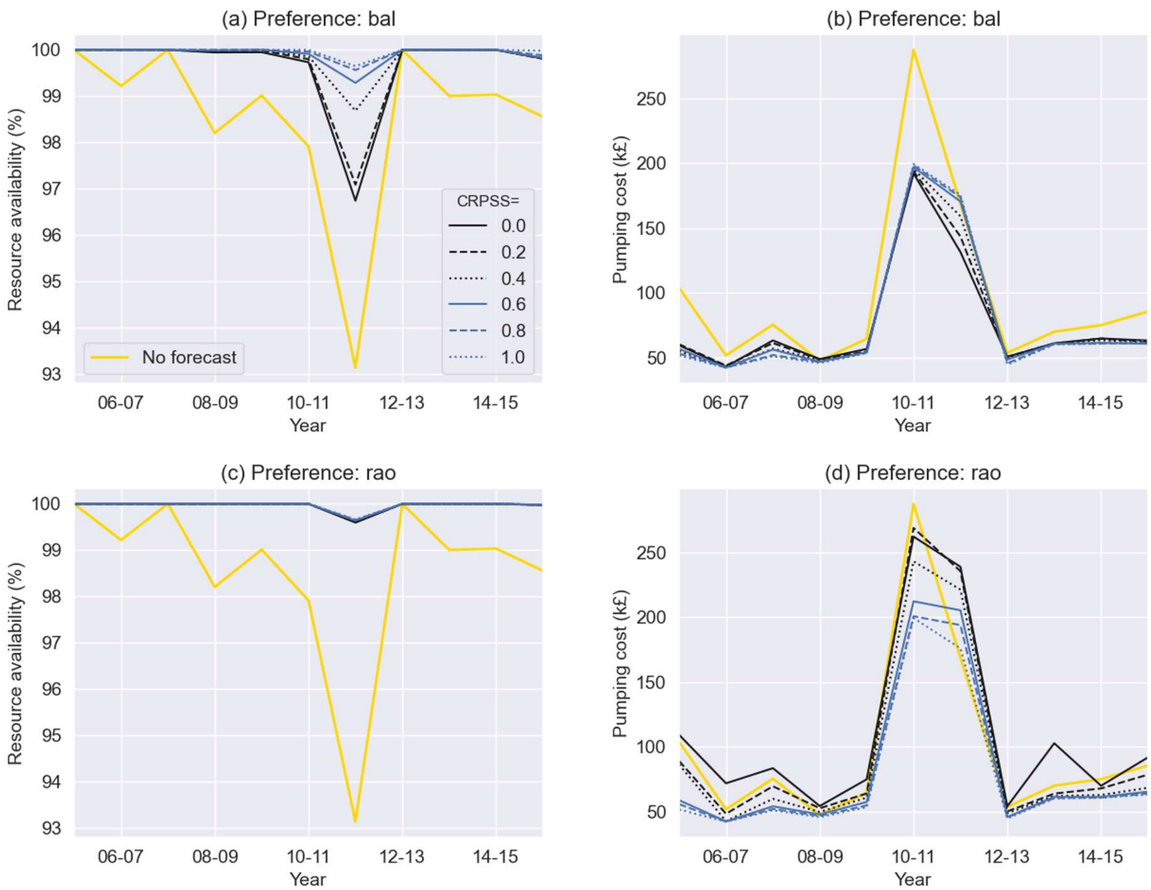

Figure 5: Simulated resource availability (RA) and pumping costs (PC) over the simulation period (from the winter 2005-06 to winter 2015-16) under the *bal* and *rao* operational preferences. Black and blue lines refer to the operations informed by forecast with different skill, the yellow line refers to the operating rules that use no forecasts.

results in different years (Figure 5). For the sake of simplicity, we focus only on two operational preferences: the balanced one (*bal*, top panels) and the one prioritising resource availability (*rao*, bottom panels) which is closest to the real-world preference of the water supply operator. Results show that in both cases the gains in value are greatest in winter 2010-11 and winter 2011-12. Both winters were particularly dry and the historical operation (yellow line) led to a very low resource availability (2011-12) and high pumping costs (2010-11).

In the *bal* case (top panels), value gains with forecast skill improvements are essentially all concentrated in those two years. The ECMWF forecast issued in November 2011, as illustrated by Figure 3, overestimated the total rainfall accumulated by April. Progressive skill improvements in our synthetic forecasts reduce the magnitude of this overestimation and lead to a better appraisal of the pumping needs during the winter 2011-12. This leads to increased pumping as forecast skill increases, improving resource availability at the expense of greater pumping costs. This behavior almost single-handedly explains Figure 4 results for the *bal* operational preference.

In contrast, the *rao* preference leads to a better repartition of benefits from increased skill across years. Because it places a premium on maximising resource availability, reservoirs are entirely full for all years except 2011-12, and over 99.5% full on that year, regardless of skill (panel (c) on Figure 5). As a result, all the improvement linked with improved skill is in reducing pumping costs (panel (d)). Whilst the brunt of pumping costs still falls on the two consecutive dry years, pumping cost savings as forecast skill increases are substantial for most years, because forecast information enables the operators to reach storage targets with less pumping. The differences between *bal* and *rao* preferences clearly lead to substantially different relationships between forecast skill and value, with value produced in different circumstances and by different mechanisms.

What is more, incremental skill improvements using the forecast families methodology highlight that, with *rao* preference, the first 0.2 skill increment is the most valuable during several winters outside of the 2010-2012 period, such as 2005-06, 2006-07, 2007-08 and 2013-14 (Figure 5.d). This finding is case-specific, but intriguing as it indicates that incremental improvements to available forecasts could then be valuable. To substantiate it, one can then explore more the operational details of the ways skill improvements bring value for the *rao* preference (Figure 6). The first skill increment curtails substantially or completely the need for pumping from the river to Reservoir 1 for 5 of the 11 years (panel (a)), showing that imperfect forecasts are enough to ensure a full reservoir and avoid underestimating the total amount of water available within the reservoir system. Further pumping savings as skill increases are through a reduced use of pumps to supply water from Reservoir 1 to the demand node (panel (b)), and an increased use of Reservoir 2 to meet that demand at no cost (panel (c)). This is because the allowable releases from reservoir 2 to leave it full on April 1^{st} are higher with better information. This is true even in years where imperfect forecasts can deliver almost the same pumping saving potential as a perfect one (e.g., 2007-08). The result of this potential to deliver near-perfect operations with imperfect forecasts is an over 10% pumping cost reduction between the benchmark (CRPSS=0) and a CRPSS=0.2 for 5 of the 11 years (panel (d)). This insight, obtained by applying the forecast families methodology, is not available by comparing the benchmark forecast only with a perfect one.

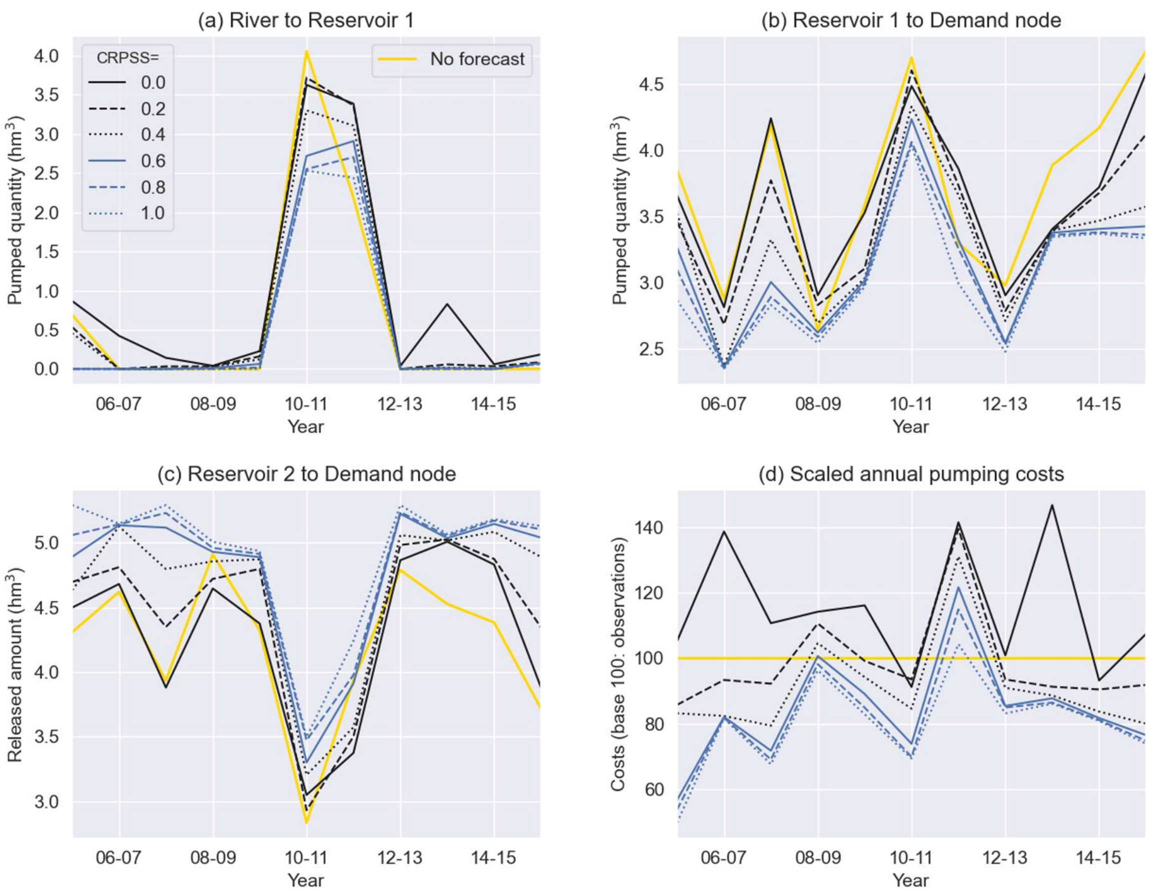

Figure 6: Simulated water fluxes over the simulation period (from winter 2005-06 to the winter 2015-16) under the *rao* operational preference. Black and blue lines refer to the operation informed by forecasts with different skill, the yellow line refers to the operating rules that use no forecasts.

Discussion

This discussion highlights a few key points from the previous section, where we illustrated the "forecast families" methodology with available ECMWF SEAS5 weather hindcasts for a water supply system in the South-West of England. The application delivered meaningful value comparisons between the forecasts generated with different skill, because these have a uniformly proportional error structure. The rest of this discussion highlights a few key points in relation to the "forecast families" methodology and its application.

Our proposed method relies on linear combinations of the hindcast with the data it is trying to predict. Because of this, it provides a first-order approximation of what an increase in skill means for a given skill score. Figure 3 illustrated this for the CRPSS, highlighting that a first-order consequence of an increase in skill is a less dispersed ensemble. It is important to note that because the linear scaling factor always verifies $k > 0$, all forecasts in a family will underestimate (or overestimate) the same events as the original hindcast. Technically, this could be fixed by picking $k < 0$, but this would affect error correlation structure if that linear scaling is not applied to all forecasts in a dataset. Likewise, applying this method does not change the timing of errors. It focuses on a single degree of freedom – error magnitude – and relates it to skill analytically. For more versatile applications, it is best to envision this method as complementary to traditional synthetic forecast generation techniques that are designed to replicate the error structure of an existing hindcast, and in particular error magnitude (e.g., Brodeur and Steinschneider, 2021). This combination would enable these traditional generators to extend the range of "what if" questions they are designed to answer..

Beyond highlighting the impact of a skill increase, our method has potential to highlight the shortcomings of widely used skill scores. For instance, it clarifies that CRPSS does not capture all aspects of forecast quality. Indeed, a common interpretation of an ensemble forecast is as an empirical probability distribution of the quantity it is trying to predict (Pappenberger et al., 2015). This means that a criterion to evaluate ensemble quality is to check whether an observation will exceed a given quantile of the ensemble with the appropriate frequency (e.g., around 50% for the ensemble median). Yet, the generation of forecast families does not affect the ranking of an observation within the ensemble used to predict it. This shows clearly that the CRPSS is not related to this ensemble quality criterion.

This work also highlighted that even though value generally increases with skill, this relationship is far from linear and is mediated by operational preferences (see Figure 4), hydrological conditions (e.g., wet/dry years as examplified in Figure 5), and the detail of the physical and socio-economic system characteristics (Figure 6). This is in line with previous studies showing that under particular conditions, an error, as identified by forecast skill scores, can still add value to the operational decisions. For instance, in certain situations within hydropower systems, recurrent positive bias (overestimation) can produce value gains (Arsenault and Côté, 2019) or losses (Cassagnole et al., 2021). This has two major consequences. First, forecast quality metrics should include an appreciation of overestimation vs. underestimation errors. Such metrics would complement usual skill scores such as the CRPSS (Peñuela et al., 2020). Second, techniques to manipulate existing forecasts such as the one presented here should be extended in the future to consider not only CRPSS, but a variety of forecast skill metrics.

This being said, the analysis presented in this paper only starts tapping into the potential insights offered by the generation of forecast families. In reality, much smaller skill increments than used in this work would best represent realistic forecast improvements, and help to assess when they can generate value. By only varying skill, our analysis also carries what can be described as a 'one-at-a-time' sensitivity analysis, whereas also varying parameters representing characteristics of the physical infrastructure (e.g. reservoir capacity), definition of the operational objectives (e.g. pumping costs) or operational preferences (e.g. the weights) would lead to a 'global' sensitivity analysis (Saltelli et al., 2008). Such global analysis would enable a more comprehensive ranking of the factors that determine forecast value, including interactions between those factors. A time-varying global sensitivity analysis (e.g., Doering et al., 2021; Rougé et al., 2021) could be used to further assess in which circumstances the different factors mostly influence value.

Generating forecast families with time-variant skill could lead to a fine-grain understanding on when skill improvements could generate value. Relating these insights with hydro-climatic processes could then help bridge the gap that often exists between those who develop forecasts and aim to improve their accuracy, e.g. meteorological agencies, and those who apply the forecasts and aim to improve their decisions, e.g. water and energy companies. This systematic analysis could then guide strategic planning and investment in forecast quality in order to obtain sufficient benefits and improved management decisions to cover the costs of improving the forecast skill (Cassagnole et al., 2021). Quality improvement can then be attained by improving the representation of targeted processes within models, and by combining model-based approaches with machine learning and deep learning techniques (Cohen et al., 2019), and with better pre-processing / bias correction techniques (Roulston and Smith, 2003; Fortin et al., 2006).

Conclusions

This paper presents a single-parameter linear scaling of forecast error to generate families of synthetic forecasts with the desired skill improvements on existing hindcasts. This "forecast families" methodology enables an exploration on the relationship between skill and value based on forecasts that resemble existing hindcasts products, instead of following an idealized probability distribution. It applies to both deterministic and ensemble forecasts, and thanks to its simplicity, it provides easily interpretable results. The application to a UK water system demonstrates the ability of the forecast families to explore the relationship between skill and value with a set of forecasts with different skill but strictly proportional error structure. This like-for-like comparison enables the most thorough exploration to date of the relationship between skill and value, using synthetic forecasts that preserve error correlation structure.

The "forecast families" method can help to explore the relationship between skill and value, and help to focus efforts in skill improvement, beyond the application to hydro-meteorological forecasts. Indeed, this is a technique that only depends on the availability of a hindcast for past data, no matter its nature. In water systems alone, it could be extended to forecasts of other variables such as water demands or energy prices – be it as a hydropower producer or an electricity consumer. Combining our method with time-variant sensitivity analysis techniques could highlight what improvements to any of these existing forecasts would yield most value. These improvements could then be achieved through the current explosion of data-driven analytics to complement and improve models' predictive ability, as well as the continuous improvement of modelling capabilities themselves.

Data availability statement

ECMWF hindcast are available under a range of licences (Vitart et al., 2017), for more information please visit https://apps.ecmwf.int/datasets/data/ s2s-reforecasts-instantaneous-accum-ecmf/levtype=sfc/type=cf/. The code used for pre-processing ECMWF hindcast and implementing the simulation-optimisation methodology is available at https://ironstoolbox.github.io/ (Peñuela et al., 2021). The reservoir system data of the case study are property of Wessex Water Ltd. and as such cannot be shared by the authors. The code for generating forecast families is available in the Zenodo open-access repository at https://doi.org/10.5281/zenodo.7327755 along with a demonstration including Figures in the paper illustrating forecast family generation.

Acknowledgments

Charles Rougé is partially funded by the UK National Environmental Research Council (NERC) via a UK Climate Resilience Embedded Researcher grant [NE/V010239/1]. Andres Peñuela is funded by the Spanish Ministry of Science and Innovation, the Spanish State Research Agency, through the Severo Ochoa and María de Maeztu Program for Centers and Units of Excellence in Research and Development [CEX2019-000968-M]. Francesca Pianosi is partially funded by the UK Engineering and Physical Sciences Research Council (EPSRC) via an Early Career Fellowship [EP/R007330/1]. For the purpose of open access, authors have applied a Creative Commons Attribution (CC BY) licence to any Author Accepted Manuscript version arising.

The authors are also very grateful to Wessex Water for the data provided. The authors wish to thank the Copernicus Climate Change and Atmosphere Monitoring Services for providing the seasonal forecasts generated by the ECMWF seasonal forecasting systems (SEAS5). Neither the European Commission nor ECMWF is responsible for any use that may be made of the Copernicus information or data it contains.

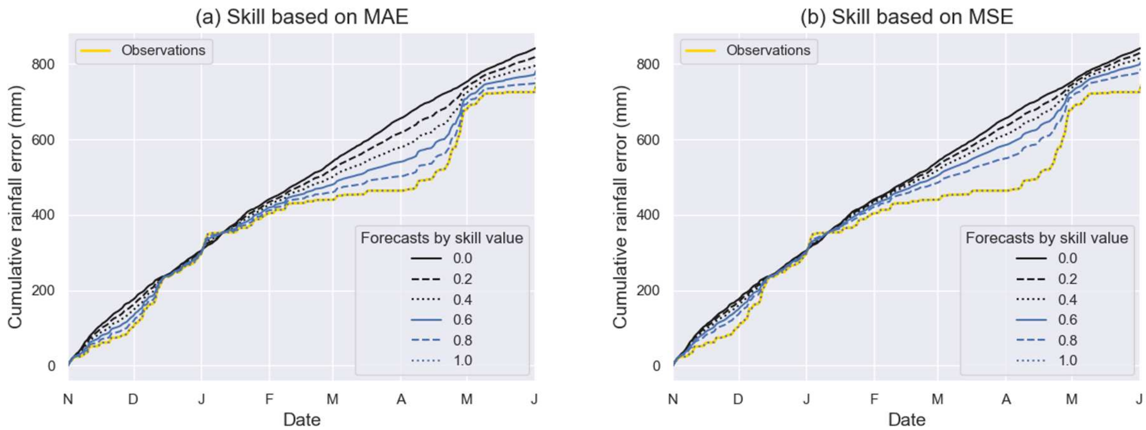

Figure A.1: Deterministic forecast families using as benchmark the ensemble mean of the bias-corrected ECMWF forecast, based on the rainfall forecast from 11/01/2011.

A Illustrating deterministic forecast families

To understand how the forecast family method applies to deterministic forecasts, one can take the ensemble mean of the same forecast. Figure A.1 illustrates this method for the two skill measures for deterministic forecasts considered in this paper, based on the mean absolute error (MAE; panel (a)) and mean squared error (MSE; panel (b)). The MAE-based family rescales error linearly as skill increases, as indicated by equation (8) with $\alpha = 1$, and demonstrating that when basing skill on the MAE, skill improvements quantify the way the average error is reduced. The MSE-based family, however, shows skill improvements with lower error reductions throughout the forecast, because of the quadratic relationship between skill and the linear error scaling factor k (equation (8) in the case $\alpha = 2$). The comparison of the two figures shows the importance of the choice of the forecast performance measure the skill is based on.

References

Alley, R.B., Emanuel, K.A., Zhang, F., 2019. Advances in weather prediction. Science 363, 342–344. https://science.sciencemag.org/content/363/6425/342.full.pdf.

Anghileri, D., Monhart, S., Zhou, C., Bogner, K., Castelletti, A., Burlando, P., Zappa, M., 2019. The value of subseasonal hydrometeorological forecasts to hydropower operations: How much does preprocessing matter? Water Resources Research n/a.

Anghileri, D., Voisin, N., Castelletti, A., Pianosi, F., Nijssen, B., Lettenmaier, D.P., 2016. Value of long-term streamflow forecasts to reservoir operations for water supply in snow-dominated river catchments. Water Resources Research 52, 4209–4225.

Arsenault, R., Côté, P., 2019. Analysis of the effects of biases in ensemble streamflow prediction (esp) forecasts on electricity production in hydropower reservoir management. Hydrology and Earth System Sciences 23, 2735–2750.

Bauer, P., Thorpe, A., Brunet, G., 2015. The quiet revolution of numerical weather prediction. Nature 525, 47–55.

Bergström, S., 1995. The HBV model, in: Singh, V. (Ed.), Computer Models of Watershed Hydrology. Water Resources Publications, p. 443–476.

Blank, J., Deb, K., 2020. pymoo: Multi-objective optimization in python. IEEE Access 8, 89497–89509.

Brodeur, Z.P., Steinschneider, S., 2021. A multivariate approach to generate synthetic short-to-medium range hydro-meteorological forecasts across locations, variables, and lead times. Water Resources Research 57, e2020WR029453. E2020WR029453 2020WR029453.

Bruno Soares, M., Alexander, M., Dessai, S., 2018. Sectoral use of climate information in Europe: A synoptic overview. Climate Services 9, 5–20.

Bruno Soares, M., Dessai, S., 2016. Barriers and enablers to the use of seasonal climate forecasts amongst organisations in Europe. Climatic Change 137, 89–103.

Cassagnole, M., Ramos, M.H., Zalachori, I., Thirel, G., Garçon, R., Gailhard, J., Ouillon, T., 2021. Impact of the quality of hydrological forecasts on the management and revenue of hydroelectric reservoirs – a conceptual approach. Hydrology and Earth System Sciences 25, 1033–1052.

Cohen, J., Coumou, D., Hwang, J., Mackey, L., Orenstein, P., Totz, S., Tziperman, E., 2019. S2s reboot: An argument for greater inclusion of machine learning in subseasonal to seasonal forecasts. WIREs Climate Change 10, e00567.

Crochemore, L., Ramos, M.H., Pechlivanidis, I.G., 2020. Can continental models convey useful seasonal hydrologic information at the catchment scale? Water Resources Research 56, e2019WR025700. E2019WR025700 2019WR025700.

Doering, K., Quinn, J., Reed, P.M., Steinschneider, S., 2021. Diagnosing the time-varying value of forecasts in multiobjective reservoir control. Journal of Water Resources Planning and Management 147, 04021031.

Ficchì, A., Raso, L., Dorchies, D., Pianosi, F., Malaterre, P.O., Overloop, P.J.V., Jay-Allemand, M., 2016. Optimal operation of the multireservoir system in the seine river basin using deterministic and ensemble forecasts. Journal of Water Resources Planning and Management 142, 05015005.

Fleming, S.W., Garen, D.C., Goodbody, A.G., McCarthy, C.S., Landers, L.C., 2021. Assessing the new natural resources conservation service water supply forecast model for the american west: A challenging test of explainable, automated, ensemble artificial intelligence. Journal of Hydrology 602, 126782.

Fortin, V., Favre, A.c., Saïd, M., 2006. Probabilistic forecasting from ensemble prediction systems: Improving upon the best-member method by using a different weight and dressing kernel for each member. Quarterly Journal of the Royal Meteorological Society 132, 1349–1369.

Giuliani, M., Crochemore, L., Pechlivanidis, I., Castelletti, A., 2020. From skill to value: isolating the influence of end user behavior on seasonal forecast assessment. Hydrology and Earth System Sciences 24, 5891–5902.

Hersbach, H., 2000. Decomposition of the continuous ranked probability score for ensemble prediction systems. Weather and Forecasting 15, 559 – 570.

Jackson-Blake, L.A., Clayer, F., de Eyto, E., French, A.S., Frías, M.D., Mercado-Bettín, D., Moore, T., Puértolas, L., Poole, R., Rinke, K., Shikhani, M., van der Linden, L., Marcé, R., 2022. Opportunities for seasonal forecasting to support water management outside the tropics. Hydrology and Earth System Sciences 26, 1389–1406.

Johnson, S.J., Stockdale, T.N., Ferranti, L., Balmaseda, M.A., Molteni, F., Magnusson, L., Tietsche, S., Decremer, D., Weisheimer, A., Balsamo, G., Keeley, S.P.E., Mogensen, K., Zuo, H., Monge-Sanz, B.M., 2019. Seas5: the new ecmwf seasonal forecast system. Geoscientific Model Development 12, 1087–1117.

Lamontagne, J.R., Stedinger, J.R., 2018. Generating synthetic streamflow forecasts with specified precision. Journal of Water Resources Planning and Management 144, 04018007.

Maurer, E.P., Lettenmaier, D.P., 2004. Potential effects of long-lead hydrologic predictability on missouri river main-stem reservoirs. Journal of Climate 17, 174 – 186.

Nayak, M.A., Herman, J.D., Steinschneider, S., 2018. Balancing flood risk and water supply in california: Policy search integrating short-term forecast ensembles with conjunctive use. Water Resources Research 54, 7557–7576.

Pappenberger, F., Ramos, M., Cloke, H., 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.

Peñuela, A., Hutton, C., Pianosi, F., 2020. Assessing the value of seasonal hydrological forecasts for improving water resource management: insights from a pilot application in the uk. Hydrology and Earth System Sciences 24, 6059–6073.

Pereira, M.V.F., Oliveira, G.C., Costa, C.C.G., Kelman, J., 1984. Stochastic streamflow models for hydroelectric systems. Water Resources Research 20, 379–390.

Peñuela, A., Hutton, C., Pianosi, F., 2021. An open-source package with interactive jupyter notebooks to enhance the accessibility of reservoir operations simulation and optimisation. Environmental Modelling & Software 145, 105188.

Rougé, C., Reed, P.M., Grogan, D.S., Zuidema, S., Prusevich, A., Glidden, S., Lamontagne, J.R., Lammers, R.B., 2021. Coordination and control − limits in standard representations of multi-reservoir operations in hydrological modeling. Hydrology and Earth System Sciences 25, 1365–1388.

Roulston, M., Smith, L., 2003. Combining dynamical and statistical ensembles. Tellus A: Dynamic Meteorology and Oceanography 55, 16–30.

Saltelli, A., Ratto, M., Andres, T., Campolongo, F., Cariboni, J., Gatelli, D., Saisana, M., Tarantola, S., 2008. Global Sensitivity Analysis: The Primer. Wiley.

Sankarasubramanian, A., Lall, U., Souza Filho, F.A., Sharma, A., 2009. Improved water allocation utilizing probabilistic climate forecasts: Short-term water contracts in a risk management framework. Water Resources Research 45.

Slater, L.J., Villarini, G., 2018. Enhancing the predictability of seasonal streamflow with a statistical-dynamical approach. Geophysical Research Letters 45, 6504–6513.

Stockdale, T., Alonso-Balmaseda, M., Johnson, S., Ferranti, L., Molteni, F., Magnusson, L., Tietsche, S., Vitart, F., Decremer, D., Weisheimer, A., Roberts, C.D., Balsamo, G., Keeley, S., Mogensen, K., Zuo, H., Mayer, M., Monge-Sanz, B., 2018. SEAS5 and the future evolution of the long-range forecast system. Technical Report 835. ECMWF.

Turner, S.W.D., Bennett, J.C., Robertson, D.E., Galelli, S., 2017. Complex relationship between seasonal streamflow forecast skill and value in reservoir operations. Hydrology and Earth System Sciences 21, 4841–4859.

Turner, S.W.D., Xu, W., Voisin, N., 2020. Inferred inflow forecast horizons guiding reservoir release decisions across the united states. Hydrology and Earth System Sciences 24, 1275–1291.

Vitart, F., Ardilouze, C., Bonet, A., Brookshaw, A., Chen, M., Codorean, C., Déqué, M., Ferranti, L., Fucile, E., Fuentes, M., Hendon, H., Hodgson, J., Kang, H.S., Kumar, A., Lin, H., Liu, G., Liu, X., Malguzzi, P., Mallas, I., Manoussakis, M., Mastrangelo, D., MacLachlan, C., McLean, P., Minami, A., Mladek, R., Nakazawa, T., Najm, S., Nie, Y., Rixen, M., Robertson, A.W., Ruti, P., Sun, C., Takaya, Y., Tolstykh, M., Venuti, F., Waliser, D., Woolnough, S., Wu, T., Won, D.J., Xiao, H., Zaripov, R., Zhang, L., 2017. The subseasonal to seasonal (s2s) prediction project database. Bulletin of the American Meteorological Society 98, 163 – 173.

Whateley, S., Palmer, R.N., Brown, C., 2015. Seasonal hydroclimatic forecasts as innovations and the challenges of adoption by water managers. Journal of Water Resources Planning and Management 141, 04014071.

Yang, G., Guo, S., Liu, P., Block, P., 2020. Integration and evaluation of forecast-informed multiobjective reservoir operations. Journal of Water Resources Planning and Management 146, 04020038.

Zhao, T., Cai, X., Yang, D., 2011. Effect of streamflow forecast uncertainty on real-time reservoir operation. Advances in Water Resources 34, 495 – 504.

Zhao, T., Zhao, J., 2014. Forecast-skill-based simulation of streamflow forecasts. Advances in Water Resources 71, 55 – 64.

Zhao, T., Zhao, J., Yang, D., Wang, H., 2013. Generalized martingale model of the uncertainty evolution of streamflow forecasts. Advances in Water Resources 57, 41 – 51.