An efficient approach for estimating streamflow forecast skill elasticity

Article

Arnal, L., Wood, A. W., Stephens, E. ORCID: https://orcid.org/0000-0002-5439-7563, Cloke, H. L. ORCID: https://orcid.org/0000-0002-1472-868X and Pappenberger, F. (2017) An efficient approach for estimating streamflow forecast skill elasticity. Journal of Hydrometeorology, 18 (6). pp. 1715-1729. ISSN 1525-7541 doi: https://doi.org/10.1175/JHM-D-16-0259.1 Available at https://centaur.reading.ac.uk/69571/

It is advisable to refer to the publisher's version if you intend to cite from the work. See Guidance on citing.

To link to this article DOI: http://dx.doi.org/10.1175/JHM-D-16-0259.1

www.reading.ac.uk/centaur

CentAUR

An Efficient Approach for Estimating Streamflow Forecast Skill Elasticity

Louise Arnal

*Department of Geography and Environmental Science, University of Reading, and European
Centre for Medium-Range Weather Forecasts, Shinfield Park,
Reading, United Kingdom*

Andrew W. Wood

Research Applications Laboratory, NCAR, Boulder, Colorado

Elisabeth Stephens

*Department of Geography and Environmental Science, University of Reading, Reading,
United Kingdom*

Hannah L. Cloke

*Department of Geography and Environmental Science, and Department of Meteorology, University
of Reading, Reading, United Kingdom*

Florian Pappenberger

*European Centre for Medium-Range Weather Forecasts, Shinfield Park, Reading, and School of
Geographical Sciences, University of Bristol, Bristol, United Kingdom*

ABSTRACT

Seasonal streamflow prediction skill can derive from catchment initial hydrological conditions (IHCs) and from the future seasonal climate forecasts (SCFs) used to produce the hydrological forecasts. Although much effort has gone into producing state-of-the-art seasonal streamflow forecasts from improving IHCs and SCFs, these developments are expensive and time consuming and the forecasting skill is still limited in most parts of the world. Hence, sensitivity analyses are crucial to funnel the resources into useful modeling and forecasting developments. It is in this context that a sensitivity analysis technique, the variational ensemble streamflow prediction assessment (VESPA) approach, was recently introduced. VESPA can be used to quantify the expected improvements in seasonal streamflow forecast skill as a result of realistic improvements in its predictability sources (i.e., the IHCs and the SCFs)—termed "skill elasticity"—and to indicate where efforts should be targeted. The VESPA approach is, however, computationally expensive, relying on multiple hindcasts having varying levels of skill in IHCs and SCFs. This paper presents two approximations of the approach that are computationally inexpensive alternatives. These new methods were tested against the original VESPA results using 30 years of ensemble hindcasts for 18 catchments of the contiguous United States. The results suggest that one of the methods, end point blending, is an effective alternative for estimating the forecast skill elasticities yielded by the VESPA approach. The results also highlight the importance of the choice of verification score for a goal-oriented sensitivity analysis.

---

1. Introduction

Unprecedented increases in computer capabilities have shaped the last several decades' advances in numerical weather prediction (NWP), and with them, the development of environmental forecasting and modeling systems. This has led to a shift in the strategy of operational forecasting centers toward more integrated modeling and forecasting approaches, such as coupled systems and Earth system models (ESMs), with the final aim to extend the limits of predictability (i.e., from subseasonal to seasonal forecasting). These developments are supported by the assimilation of more and better-quality observation data as well as the increase in model resolutions and complexity. However, such advances can be very expensive and data hungry and may not yield proportional improvements.

Seasonal hydrological forecasts are predictions of the future states of the land surface hydrology (e.g., streamflow), up to a few months ahead. They are valuable for applications such as reservoir management for hydropower, agriculture and urban water supply, spring flood and drought prediction, and navigation, among others (Clark et al. 2001; Hamlet et al. 2002; Chiew et al. 2003; Wood and Lettenmaier 2006; Regonda et al. 2006; Luo and Wood 2007; Kwon et al. 2009; Cherry et al. 2005; Viel et al. 2016). They have the potential to provide early warning for increased preparedness (Yuan et al. 2015). Traditionally, seasonal streamflow forecasts have relied upon land surface memory, the persistence in the land surface (e.g., catchment) initial hydrological conditions (IHCs; of soil moisture, groundwater, snowpack, and the current streamflow). IHCs are one of the most important predictability sources of seasonal streamflow forecasts and were thus the starting point for the development of the ensemble streamflow prediction (ESP) approach in the 1970s (Wood et al. 2016b). The ESP was first developed and used for reservoir management purposes. It is produced by running a hydrological model with observed meteorological inputs to produce current observed IHCs, from which the forecast is started, and the forcing over the forecast period is undertaken using an ensemble of historical meteorological observations (Day 1985). The ESP method assumes that the model states to initialize a forecast are perfectly estimated, while the future climate is completely unknown. However, the skill of the ESP decreases significantly after one to a few months of lead time over most parts of the world because of a decrease in the land surface memory with time. The achievable predictability from the ESP thus depends on the persistence of the IHCs, which can vary as a function of the season (i.e., the transition between dry and wet seasons can, for example, be hard to forecast) and the location and size of the catchment (i.e., the streamflow in a large catchment with a slow response time and/or situated in a region with negligible precipitation inputs during the forecast period will for example be easier to forecast; Wood and Lettenmaier 2008; Shukla et al. 2013; van Dijk et al. 2013; Yuan et al. 2015).

More recently, seasonal climate predictability derived from large-scale climate precursors [e.g., El Niño–Southern Oscillation (ENSO) and the North Atlantic Oscillation (NAO)] has been used to enhance seasonal streamflow forecasting (e.g., Wood et al. 2002; Yuan et al. 2013; Demargne et al. 2014; Mendoza et al. 2017). Such systems produce streamflow forecasts by initializing a hydrological model to estimate IHCs and forcing the model with inputs based on seasonal climate forecasts (SCFs; of temperature and precipitation) instead of historical observations. Their skill is also still limited because of the rapid decrease in precipitation forecasting skill beyond two weeks of lead time, and the skill is variable in both space and time (Yuan et al. 2011; van Dijk et al. 2013; Slater et al. 2017). In Europe, for instance, the skill is higher in winter in regions where the winter precipitation is highly correlated with the NAO. Regions with high skill include the Iberian Peninsula, Scandinavia, and regions around the Black Sea (Bierkens and van Beek 2009). In the contiguous United States (CONUS), the skill is on average higher over (semi)arid western catchments, due to the persistence of the IHCs influence up to three months of lead time. The skill can be higher in some regions of the western CONUS (i.e., California, the Pacific Northwest, and Great Basin) in the winter and fall due to higher precipitation forecasting skill in strong ENSO phases (Wood et al. 2005).

Increasing the seasonal streamflow forecast skill remains a challenge: one that is being tackled by improving IHCs and SCFs using a variety of techniques. Techniques include model developments and data assimilation and can vary in computational expense. However, over the past several decades, it has been shown that operational streamflow forecast quality has not significantly improved (Pagano et al. 2004; Welles et al. 2007). This is the motivation for the use of sensitivity analysis techniques to guide future forecasting developments for seasonal streamflow forecasting and is the basis for this paper.

It is in this context that the attribution of seasonal streamflow forecast uncertainty to the IHC and SCF errors has been researched extensively. Wood and Lettenmaier (2008) introduced a method based on two hindcasting end points: the ESP and the reverse ESP. In contrast to the ESP, which only represents the

uncertainty in the future climate, the reverse ESP only represents the uncertainty in IHCs by using an ensemble of initial model states taken from historical simulations to initialize a prediction forced by a single set of observed meteorological inputs. Typically, the input uncertainty attenuates over a period of months as the influence of the perfect future climate input increasingly determines model states.

Comparing the skill of the ESP versus reverse-ESP seasonal streamflow forecasts allows one to identify the dominant predictability source (and conversely uncertainty source) of seasonal streamflow forecasting (i.e., the IHCs or the SCFs), and its evolution in both space and time. It was successfully used to disentangle the relative importance of initial conditions and boundary forcing errors on seasonal streamflow forecast uncertainties by several authors: for example, for catchments in the United States (Wood and Lettenmaier 2008; Li et al. 2009; Shukla and Lettenmaier 2011), in France (Singla et al. 2012), in Switzerland (Staudinger and Seibert 2014), in China (Yuan et al. 2016; Yuan 2016), and in the Amazon (Paiva et al. 2012), as well as for the entire globe (Shukla et al. 2013; Yossef et al. 2013; MacLeod et al. 2016). This work is instructive as it enables the dominant predictability source to be identified (i.e., where efforts and resources should be targeted) to focus improvement, which could potentially lead to more skillful seasonal streamflow predictions.

This method was extended by Wood et al. (2016a, hereafter W16) via a method called variational ensemble streamflow prediction assessment (VESPA), which involves assessing intermediate IHC and SCF uncertainty points between the perfect and climatological points applied in ESP and reverse ESP. The approach allows the calculation of a metric called "skill elasticity," that is, the sensitivity of streamflow forecast skill to IHC and SCF skill changes. A key drawback of the VESPA approach, however, is that it is computationally intensive. For each catchment and initialization month of a forecast, the response surface was defined through the use of dozens of multidecadal variable-skill ensemble hindcasts, ultimately amounting to millions of simulations. In contrast, the ESP and reverse-ESP skill can be estimated from a single set of ensemble hindcasts spanning a historical period. The IHC and SCF skill variation method was also highly specific to the particular model state configuration and involved a relatively simplistic linear blending procedure. The elasticity calculations were furthermore based only on a single verification score of forecast skill (i.e., coefficient of determination R^2) for the analysis. An ensemble forecast has many attributes, for example, the skill, the reliability, the resolution, and the uncertainty of the forecast, among others. To obtain a complete picture of the forecast quality, the scores should encompass many of these attributes, as each verification score will give us different information about the forecast quality.

The drawbacks of VESPA motivate us to assess two computationally inexpensive methods of estimating the forecast skill elasticities, using only the original ESP and reverse-ESP results that depend on the single hindcast series as mentioned above. The two methods are termed end point interpolation (EPI) and end point blending (EPB). In the first part of this paper, we compare results from the two methods tested on 18 catchments of the CONUS to the original results from the VESPA, using a single verification score. The objective of this part is to investigate whether the new methods can discriminate the influence of IHC and SCF errors on seasonal streamflow forecasting uncertainties and to assess the ability of those new methods to correctly estimate the forecast skill elasticities. In the second part, additional verification scores are applied for streamflow forecast verification, supporting the second objective of the paper, which is to explore the sensitivity of the results obtained from the two new methods and the VESPA approach to the choice of the verification score.

2. Methods, data, and evaluation strategy

a. The VESPA approach

In this work, as in W16, the term "perfect" refers to current observed meteorological data and the term climatological refers to the whole distribution of historical observed data. Figure 1 presents the ESP (Fig. 1a), the reverse ESP (Fig. 1b), the climatology (Fig. 1c), and the VESPA forecast (Fig. 1d), as generated in W16. The ESP, the reverse ESP, the perfect forecast, and the climatology are all end points of the uncertainty in the sense that the uncertainty in those forecasts is either perfect or climatological. They are the end points of the VESPA approach.

VESPA aims to produce streamflow forecasts from IHCs and SCFs with an uncertainty situated between the perfect and the climatological uncertainty (Fig. 1d). Forecasts were generated by linearly blending the climatological and perfect IHCs (i.e., model moisture states) and the climatological and perfect SCFs (i.e., meteorological forcings of precipitation, evapotranspiration, and temperature), subsequently used to run the hydrological model. The weights used for blending the data were (w = 0, 0.05, 0.10, 0.25, 0.50, 0.75, 0.90, 0.95, 1.0), applied so that a weight of zero is the perfect knowledge and unity is the climatological knowledge, with w_{IHC} and w_{SCF} denoting the weights used to blend the IHCs and the SCFs, respectively (W16). An ESP forecast results from the weights $w_{IHC} = 0$ and $w_{SCF} = 1$, the reverse ESP from $w_{IHC} = 1$ and $w_{SCF} = 0$, the perfect
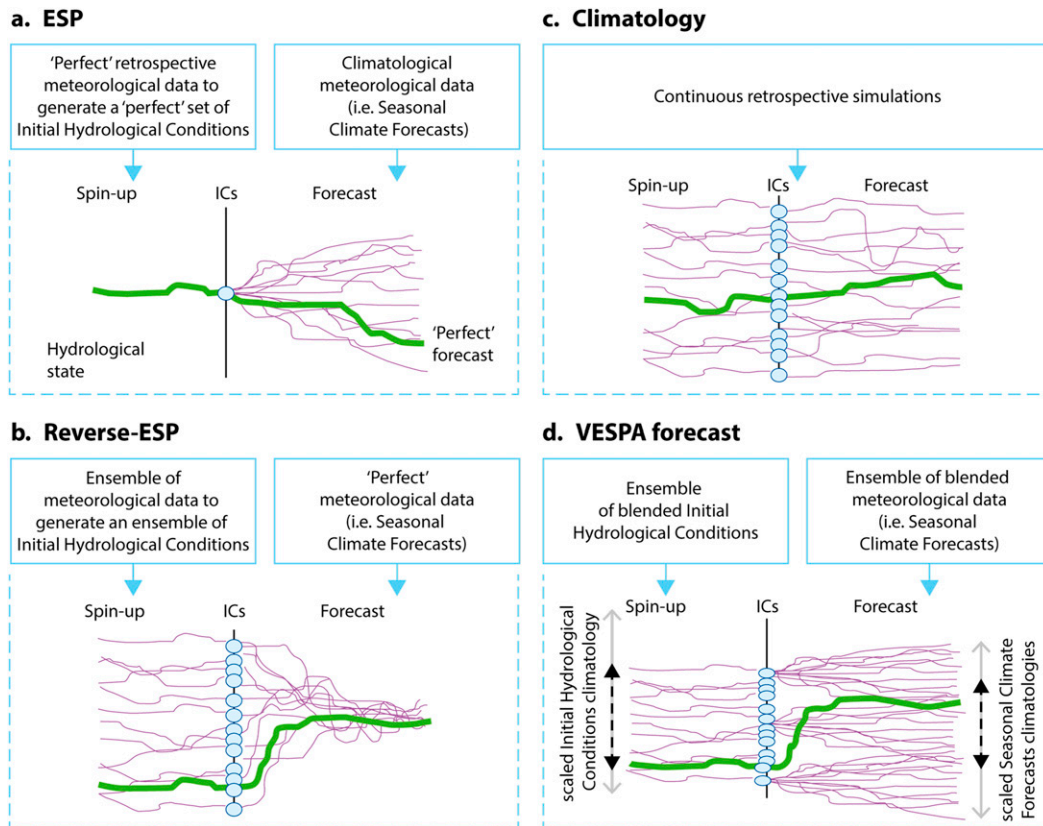
FIG. 1. Schematic of (a) the ESP, (b) the reverse ESP, (c) the climatology, and (d) the VESPA (this figure is adapted from Fig. 3 in W16).

forecast from $w_{IHC} = 0$ and $w_{SCF} = 0$, and the climatology from $w_{IHC} = 1$ and $w_{SCF} = 1$.

To plot the skill of the VESPA forecasts as a function of the IHC and SCF skill, W16 used skill surface plots (Fig. 2), interpolating forecast skill results from different IHC and SCF weighting combinations. The axes represent the SCF and IHC skill, derived respectively from the blending weights w_{SCF} and w_{IHC} using the following equations (W16):

$$SCF\ skill = 100 \times (1 - w_{SCF}^2) \quad \text{and} \tag{1}$$

$$IHC\ skill = 100 \times (1 - w_{IHC}^2). \tag{2}$$

The SCF and the IHC skill values obtained from these equations are the percentage of climatological variance explained in the respective predictability source (i.e., SCF and IHC; W16). Each SCF skill–IHC skill combination corresponds to a specific VESPA forecast, the skill of which can be plotted on the skill surface plot (black plus signs in Fig. 2). The blue circles are the end points of the VESPA forecasts: the reverse ESP (revESP in Fig. 2), the perfect forecasts, the ESP, and the climatology (climo in Fig. 2). The skill surface plots are hence a graphical representation of the

response surface obtained from the VESPA sensitivity analysis.

The VESPA seasonal streamflow forecasts were generated by W16 using lumped Sacramento Soil Moisture Accounting (SAC-SMA) and SNOW-17 catchment models for unimpaired catchments. The models were forced with daily inputs in precipitation, temperature, and potential evapotranspiration and were calibrated and validated against observed daily streamflow from the U.S. Geological Survey (USGS). Eighty-one skill variations of a 30-yr hindcast (from 1981 to 2010) were produced for 424 catchments in the CONUS, starting at the beginning of each month (i.e., forecast initialization dates), with lead times up to 6 months.

b. Alternative methods to the VESPA approach

In this paper we present two alternative methods of the VESPA approach, the EPI and the EPB. These methods aim to reproduce the response surface obtained from the VESPA approach by using the same 30-yr hindcast ensembles produced by W16, aggregated over the first three months with zero lead time for each initialization date (referred to as 3-month streamflow forecast hereafter) and corresponding exclusively to the end points (i.e., the ESP, the reverse ESP, the perfect forecast, and the climatology).
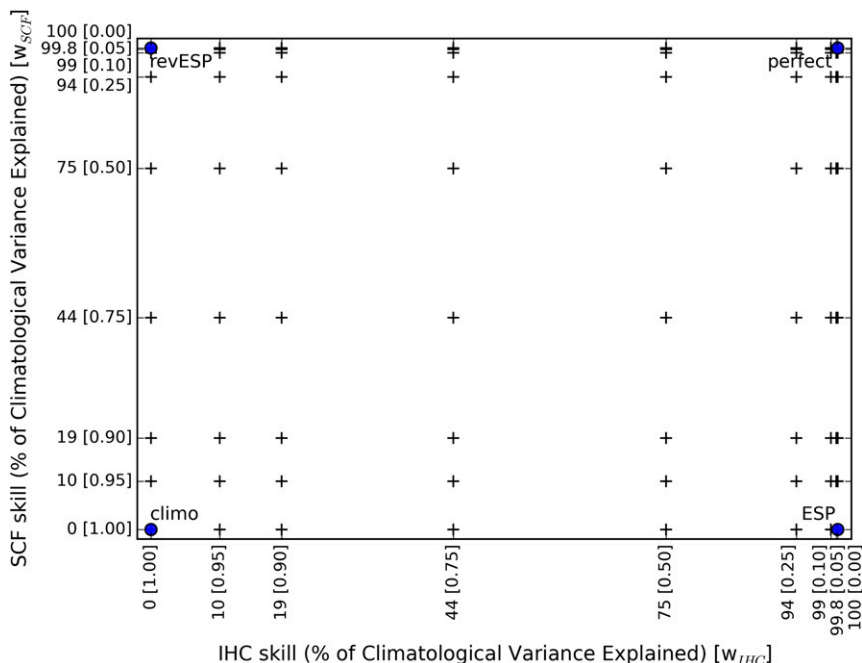
FIG. 2. Schematic of a skill surface plot. The *y* and *x* axes display the SCF and the IHC skill, respectively. They are expressed as a percentage of the climatological variance explained in the respective predictability source. The blending weights, w_{SCF} and w_{IHC}, from which the skill values are derived are shown in square brackets in the figure.

The two new methods were tested for a subset of the CONUS-wide catchment dataset presented in W16 (Fig. 3), comprising 18 catchments from the large USGS Hydro-Climatic Data Network (HCDN; Lins 2012). The 18 selected catchments cover a large range of hydrometeorological conditions, including the maritime climate regime of the U.S. West Coast catchments; the humid regime of the eastern United States (south of the Great

Lakes) with rainfall-driven runoff and variable winter snow in the most northern catchments; and the Intermountain West and northern Great Plains regions, where streamflow is greatly influenced by the snow cycle.

1) END POINT INTERPOLATION

The EPI produces a response surface by interpolating the forecast skill of the end points throughout the skill

FIG. 3. Map of the 18 catchments of the CONUS selected for the analysis and the HCDN regions (dark blue outlines).

surface plot. Both linear (i.e., linear barycentric interpolation) and cubic interpolation techniques were tested. However, results will be shown for the linear interpolation only as the cubic interpolation did not provide noticeable improvements to the linear interpolation given that the interpolation is based on only four points situated at the corners of the response surface. The linear EPI was performed for each forecast initialization date and for each catchment.

2) END POINT BLENDING

The EPB generates hindcasts for each w_{SCF}–w_{IHC} combination (i.e., each plus sign in Fig. 2; w_{SCF} and w_{IHC} are selected to be the same blending weights used by W16, for the purpose of comparison). For each combination point, a new ensemble of 100 members was generated by blending the four end points, given a specific weighted average. The percentage of each end point used [EP(%); i.e., the number of members randomly selected from each end point], was calculated for each combination point using the following equation:

$$EP(\%) = (1 - |x_{EP} - w_{IHC}|) \times (1 - |y_{EP} - w_{SCF}|), \quad (3)$$

where x_{EP} and y_{EP} are the w_{IHC} and w_{SCF} values of the end point for which the percentage is calculated, respectively. For example, if the w_{IHC} and w_{SCF} match the end point values, 100% of the EPB hindcast members are resampled from that end point (i.e., the end point skill is reproduced). This was done for each forecast initialization date and for each catchment.

To produce the skill surface plots for the EPB method, the SCF and IHC skill was calculated using the same equations as in W16 [i.e., Eqs. (1) and (2), respectively].

c. The evaluation strategy

The aim of this paper is to compare two computationally inexpensive alternative methods to the VESPA approach, the EPI and the EPB. To this end, the paper unfolds into two distinct objectives. First, we want to investigate whether the EPI and/or the EPB can discriminate the influence of IHC and SCF errors on seasonal streamflow forecasting uncertainties and reproduce VESPA skill elasticity estimates. This will validate the use of one or both methods as alternative to the VESPA approach. Second, we want to explore the sensitivity of the results obtained from the EPI, the EPB, and the VESPA methods to the choice of the verification score. This will be an attempt to demonstrate the importance of the choice of the verification score for forecast verification and communication.

1) CAN EPI AND EPB DISCRIMINATE THE INFLUENCE OF IHC AND SCF ERRORS ON SEASONAL STREAMFLOW FORECAST UNCERTAINTIES?

To explore the first objective of this paper, skill surface plots were produced for the EPI, the EPB, and the VESPA methods. As in W16, the seasonal streamflow forecast skill depicted in the skill surface plots was calculated from the R^2 of forecast ensemble means with the observations, where perfect forecasts (model simulations driven by the observed meteorology) were treated as observations to calculate the R^2. As discussed at length in W16, this choice deliberately excludes the model errors as a source of forecast uncertainty.

The skill surface plots obtained from the EPI and the EPB methods were subsequently compared qualitatively and quantitatively to the skill surface plots obtained for the VESPA approach. The qualitative analysis consisted in visually inspecting the patterns contained in the skill surface plots, giving an indication of the dominant predictability source on the streamflow forecast skill. The quantitative analysis of the results was based on the calculation of the skill elasticities for the IHCs and the SCFs (E_{IHC} and E_{SCF}, respectively), for the EPI, the EPB, and the VESPA methods, averaged across three transects of a quadrant situated in the center of the response surface, according to the following equations:

$$E_{IHC} = 100 \times \left\{ \frac{S(F[75,19]) - S(F[19,19])}{75\% - 19\%} + \frac{S(F[75,44]) - S(F[19,44])}{75\% - 19\%} + \frac{S(F[75,75]) - S(F[19,75])}{75\% - 19\%} \right\} \Big/ 3 \tag{4}$$

and

$$E_{SCF} = 100 \times \left\{ \frac{S(F[19,75]) - S(F[19,19])}{75\% - 19\%} + \frac{S(F[44,75]) - S(F[44,19])}{75\% - 19\%} + \frac{S(F[75,75]) - S(F[75,19])}{75\% - 19\%} \right\} \Big/ 3. \tag{5}$$

The numerators, expressed as $S(F[\cdot]) - S(F[\cdot])$, contain the gradients in the streamflow forecast skill between IHC skill (or SCF skill) values of 75% and 19% (the denominator). The values in between the

square brackets of the numerator are the IHC skill followed by the SCF skill values, which indicates a certain w_{SCF}–w_{IHC} combination point in the example skill surface plot in Fig. 2. In the denominator, the IHC and SCF skill gradients are gradients in the percentage of the climatological variance explained in the respective predictability source. The skill elasticities (E_{IHC} and E_{SCF}) are positively oriented, where a skill elasticity of zero is obtained when the predictability source has no influence on the skill of the streamflow forecast, while positive (negative) elasticities mean that an improvement in the predictability source will lead to higher (lower) streamflow forecast skill. The skill elasticities were calculated for all three methods and for the 3-month streamflow forecasts produced for each catchment and forecast initialization date.

The only difference between Eqs. (4) and (5) and the skill elasticities calculated in W16 is that they chose to calculate skill elasticities around the ESP point in the skill surface plots. Here, we choose to calculate skill elasticities across a quadrant within the skill surface plot (between 75% and 19% of the climatological variance explained in the IHC and the SCF) in order for the skill elasticity values calculated in this paper to reflect the forecast skill gradients within the response surface. This is done differently to W16 because the aim of this paper is to compare (qualitatively and quantitatively) the skill surface plots obtained from the EPI and the EPB methods to the VESPA approach.

2) WHAT IS THE SENSITIVITY OF THE RESPONSE SURFACE TO THE CHOICE OF THE VERIFICATION SCORE?

To investigate the second objective of this paper, several verification scores were calculated for each method (i.e., the EPI, the EPB, and the VESPA approach). These scores were selected in order to cover key attributes of the forecasts verified, and they include

- the mean absolute error (MAE) of forecast ensemble means, relative to the perfect forecasts and
- the continuous rank probability score (CRPS) and its decomposition:
  - the potential CRPS (CRPSpot), where CRPSpot = resolution − uncertainty, and
  - the reliability part of the CRPS (CRPSreli).

The potential CRPS is the CRPS value that a forecast with perfect reliability would have. The uncertainty is the variability of the observations and the resolution is the ability of the forecast to distinguish situations with distinctly different frequencies of occurrence. The CRPS reliability is a measure of the bias and the spread of the system.

The CRPS was chosen as it is a widely used score to assess the overall quality of an ensemble hydrometeorological forecast. The CRPS moreover has the advantage that it can be decomposed into different scores in order to look at the many different attributes of an ensemble forecast. The CRPS for a single forecast is equivalent to the MAE, which is why the latter was chosen.

For all of the above verification scores, the corresponding skill scores were calculated for each point of the skill surface plots from

$$\text{skill score}_{\text{forecast}} = 1 - \frac{\text{score}_{\text{forecast}}}{\text{score}_{\text{reference}}}, \qquad (6)$$

where the score$_{\text{reference}}$ is the score of the climatology point, for each method, each initialization date, and each catchment. A perfect forecast results in a forecast skill score of unity and a forecast with equal quality as the reference forecast corresponds to a skill score of zero. Any forecasts of lower quality than the reference forecast produce negative skill score values. Skill scores were calculated in order to have a similar score range as the R^2 (i.e., a climatological score of zero and a perfect score of one), for comparative purposes.

Skill elasticities were subsequently calculated for all the skill scores, using Eqs. (4) and (5), for all three methods and for the 3-month streamflow forecasts produced for each catchment and forecast initialization date. From these skill elasticity values, the influence of improvements in the IHCs and SCFs on the seasonal streamflow forecast skill can be assessed, in terms of the forecasts' overall performance (considering the mean of the ensemble or the full ensemble spread, from the MAE and the CRPS, respectively), their resolution and uncertainty (CRPSpot), and their reliability (CRPSreli).

3. Results

a. Can EPI and EPB discriminate the influence of IHC and SCF errors on seasonal streamflow forecast uncertainties?

For the first part of this study, the Crystal River (Colorado; USGS gauge 009081600), a snowmelt-driven catchment, will be used as a test case to illustrate the skill surface plots obtained from the EPI and the EPB methods, compared to the VESPA approach. Precipitation is the highest in winter and spring in this catchment and falls as snow between November and April. In April, the snow starts melting and consequently the soil moisture and streamflow both increase.

Figure 4 displays the skill surface plots obtained for the VESPA (Fig. 4a), the linear EPI (Fig. 4b), and the EPB methods (Fig. 4c), from R^2 for the 3-month streamflow forecast for the Crystal River, for initializations on the first of each month (each row in Fig. 4).
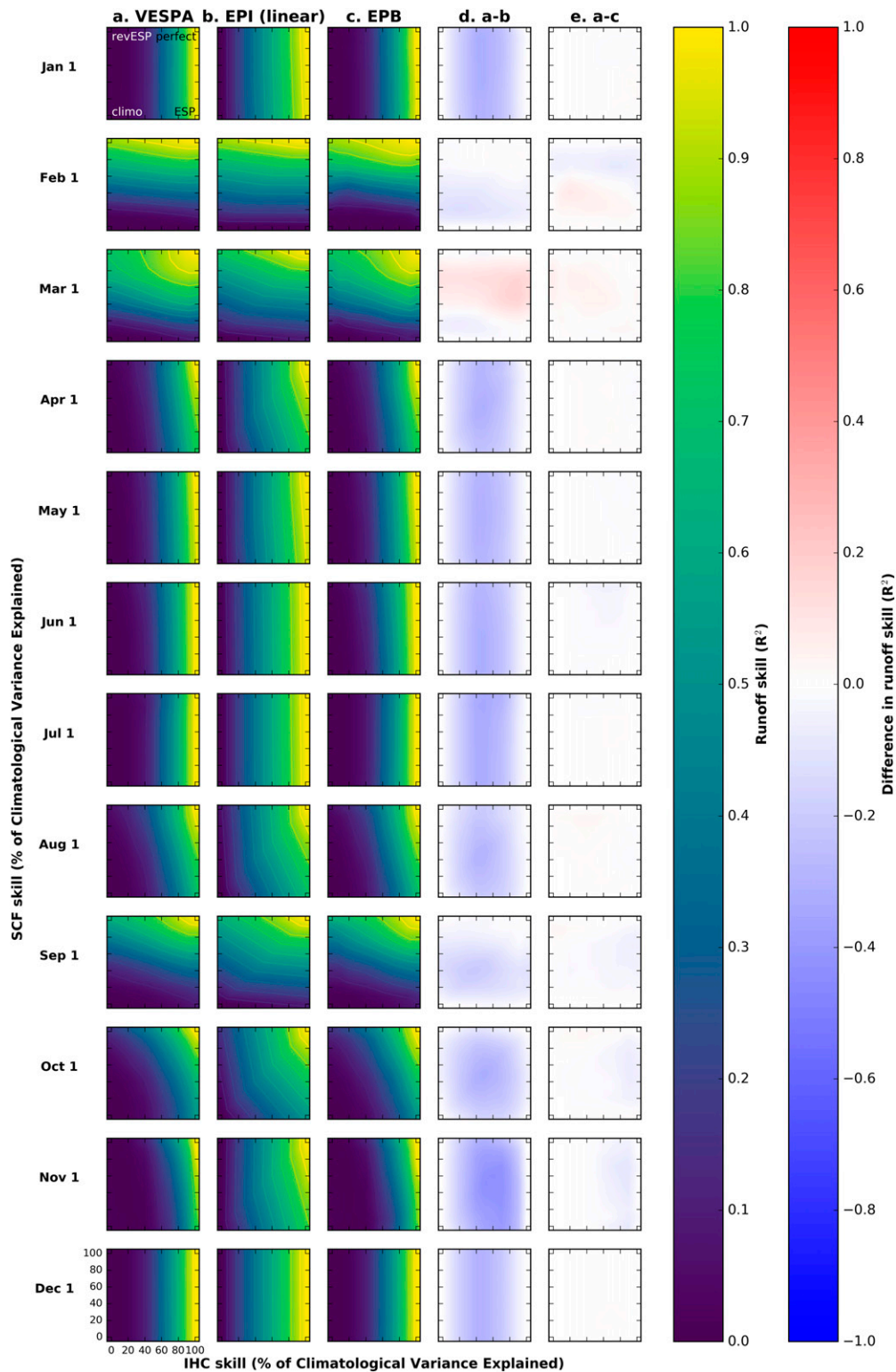
FIG. 4. Skill surface plots obtained for (a) the VESPA, (b) the linear EPI, and (c) the EPB methods. The skill is calculated from the R^2 of the 3-month streamflow forecast ensemble means against the perfect forecasts, for hindcasts produced from 1981 to 2010 for the Crystal River (USGS gauge 009081600), with forecast initializations on the first day of each month. Differences between the skill surface plots obtained for (d) the VESPA and linear EPI methods and (e) the VESPA and EPB methods are also shown.

Figures 4d and 4e show the differences between the skill surface plots obtained for the VESPA and EPI methods and the VESPA and EPB methods, respectively. A first visual comparison of the skill surface plots obtained from the linear EPI method (Fig. 4b) and the EPB method (Fig. 4c) with those obtained from the VESPA approach (Fig. 4a) for the Crystal River tells us that the skill surface plots obtained from all three methods are very similar. For each initialization date, the orientation of the gradients in streamflow forecast skill appears identical. The EPI and the EPB methods seem to correctly indicate the dominant predictability source on the 3-month streamflow forecast skill, for each initialization date for this catchment. Similar results were obtained for the other 17 catchments (see Figs. S1–S17 in the supplemental material). Forecasts made on the first of February, March, and September show a sensitivity to the SCF skill (i.e., horizontal or near to horizontal orientation of the streamflow forecast skill gradients), while all other forecasts are dominantly sensitive to the IHC skill (i.e., vertical or near to vertical orientation of the streamflow forecast skill gradients).

The gradients in streamflow forecast skill contained in the EPI skill surface plots (Fig. 4b) differ moderately from the gradients obtained from the VESPA approach (Fig. 4a). This can be observed in Fig. 4d, showing the differences between the skill surface plots obtained for both methods. The VESPA approach gives very strong gradients, causing a rapid decrease in streamflow forecast skill with a decrease in one of the predictability sources' skill, depending on the initialization date. In comparison, the EPI method indicates a gradual decrease in streamflow forecast skill with a decrease in one of the two predictability sources, depending on the initialization date. The streamflow forecast skill gradients produced by the EPI method are a reflection of the interpolation method used (i.e., here linear), and because the corner points lack information about describing curvature of the surface at interior points, they cannot fully capture nonlinearities in the skill gradients across the skill surface. For some interior points, this limitation of the EPI method could estimate very different skill elasticities than those obtained from the VESPA approach.

The skill surface plots produced by the EPB method (Fig. 4c) show minor differences in the streamflow forecast skill gradients when compared to the skill surface plots generated by the VESPA approach (Fig. 4a). This can be seen in Fig. 4e, which shows the differences between the skill surface plots obtained for both methods. To further inspect those differences, they will be explored quantitatively (i.e., by comparing the skill elasticities) below.

To quantify the accuracy of the patterns contained in the EPI and the EPB skill surface plots compared to the patterns of the VESPA skill surface plots, IHC and SCF skill elasticities (i.e., E_{IHC} and E_{SCF}, respectively) were calculated across a quadrant situated within the response surface for all three methods, for the 18 catchments and each forecast initialization date, from Eqs. (4) and (5), respectively. Figure 5 presents the skill elasticities for nine of the 18 catchments (the plots for the other nine catchments are shown in Fig. S18). Each plot corresponds to a catchment and shows the skill elasticities obtained from the VESPA, the linear EPI, and the EPB methods as a function of the forecast initialization date. From the nine different plots, the skill elasticities given by the EPB method appear almost identical to the VESPA approach, whereas the skill elasticities obtained from the EPI method differ in some places. This confirms that the patterns of the EPB method are very similar to the patterns of the VESPA approach, with it being the closest out of the two tested methods.

The value of the SCF skill elasticity (i.e., E_{SCF}) in relation to the value of the IHC skill elasticity (i.e., E_{IHC}), for a given method, indicates the dominant predictability source on the 3-month streamflow forecast skill (here calculated from the R^2). For a selected method, equal SCF and IHC skill elasticity values signifies that equal improvements in both the SCFs and the IHCs will lead to equal improvements in the streamflow forecast skill. If E_{SCF} is superior (inferior) to E_{IHC}, it reflects a larger potential increase in streamflow forecast skill by improving the SCFs (IHCs). Although the EPI method almost always indicates the same dominant predictability source as the two other methods, the degree of influence of changes in IHC and SCF skill on the streamflow forecast skill (i.e., the exact values of the skill elasticities) often differs. For many catchments and forecast initialization dates, the EPI appears to underestimate the skill elasticities produced by the VESPA method.

The nine different catchments for which the skill elasticities are presented in Fig. 5 display three different types of behavior, best captured by the VESPA approach and the EPB method. For the three catchments in Fig. 5 (left), improvements in the IHCs would yield the highest improvements in the 3-month streamflow forecast skill for spring to summer initializations (April–August for the Crystal River, March–July for the Fish River, and March–June for the Middle Branch Escanaba River) and in the winter (October–January for the Crystal River, November–December for the Fish River, and in December for the Middle Branch Escanaba River). SCF improvements would lead to better 3-month streamflow forecast skill for forecasts initialized in the late winter and summer to fall (February–March and September for the Crystal River, February and August–October for the Fish River, and January–February and July–September for the Middle Branch Escanaba River). For the three catchments in Fig. 5 (middle),
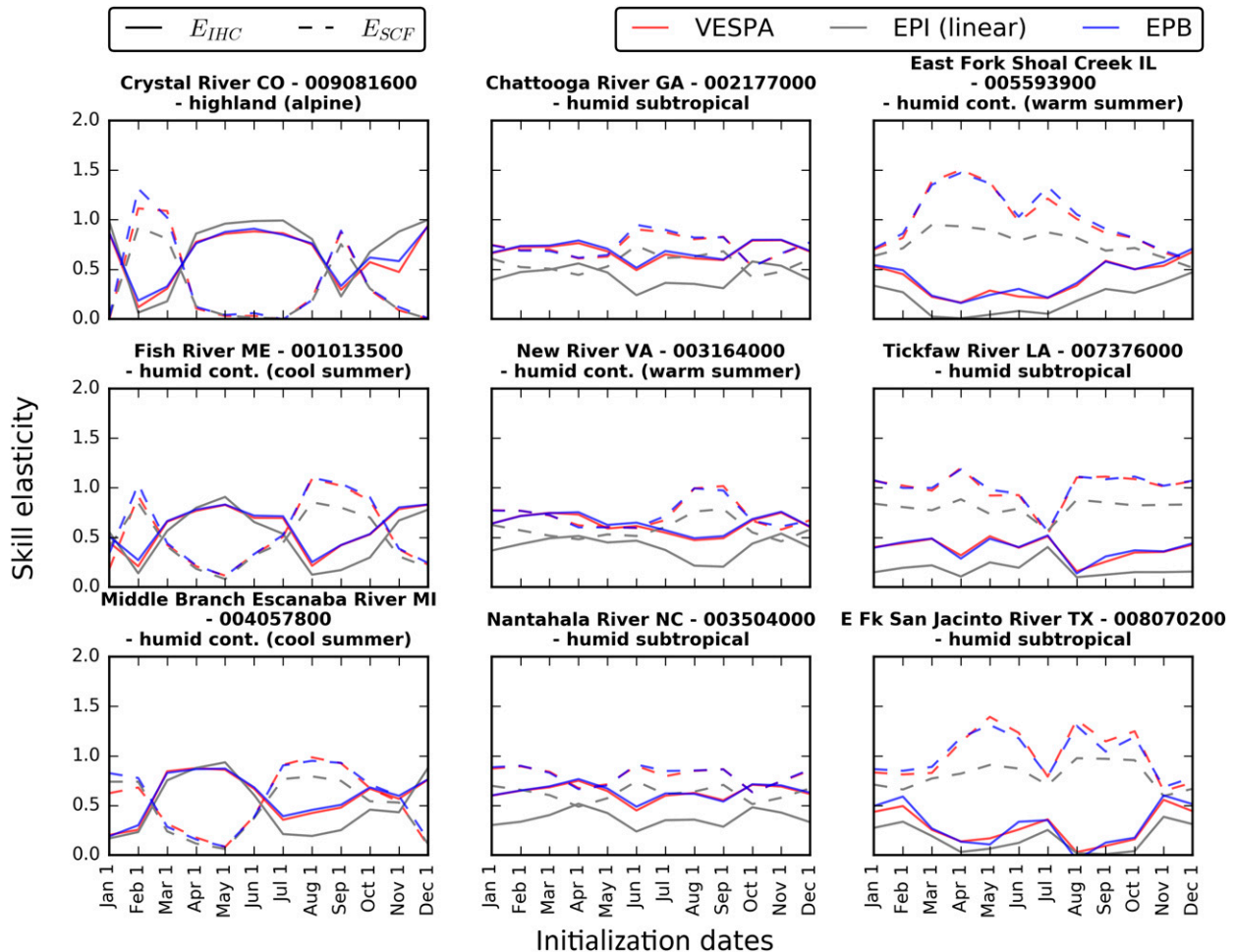
FIG. 5. Streamflow forecast skill elasticities for the IHCs (i.e., E_{IHC}, solid line) and the SCFs (i.e., E_{SCF}, dashed line), calculated across a quadrant situated within the 3-month streamflow forecast skill surface plots for the VESPA (red), the linear EPI method (gray), and the EPB method [blue; using Eqs. (4) and (5)]. Each plot shows the evolution of the IHC and SCF skill elasticities with the initialization date for a given catchment. The climatological regions of the catchments are indicated in the plots' headings. The skill surface plots from which these skill elasticities were calculated are presented in Fig. 4 and Figs. S1–S17.

a notable feature is that the 3-month streamflow forecast skill would benefit from SCF improvements for summer initializations (June–September for the Chattooga and the Nantahala Rivers and July–September for the New River). Finally, for the three catchments in Fig. 5 (right), the 3-month streamflow forecast skill would benefit from improvements in the SCFs for all initialization dates. This is true with the exception of forecasts initialized in December for East Fork Shoal Creek. It is important to note that there is uncertainty around these estimates. However, this is a good first indication of the sensitivity of 3-month streamflow forecast skill (measured from the R^2) to IHC and SCF errors for each forecast initialization date and each catchment.

The skill elasticities produced by the EPB method appear to be almost identical to the skill elasticities obtained from the VESPA approach, with occasional marginal differences. This suggests that the EPB method captures nearly exactly the degree of influence of changes in IHC and SCF skill on the streamflow forecast skill, obtained from the VESPA approach. Both methods additionally indicate the same dominant predictability source: the predictability source which, once improved, could lead to the largest increase in 3-month streamflow forecast skill. The EPB method will therefore be used as an alternative to the VESPA approach to investigate the second objective of this paper.

b. *What is the sensitivity of the response surface to the choice of the verification score?*

To investigate the sensitivity of the response surface to the choice of the verification score, and therefore to
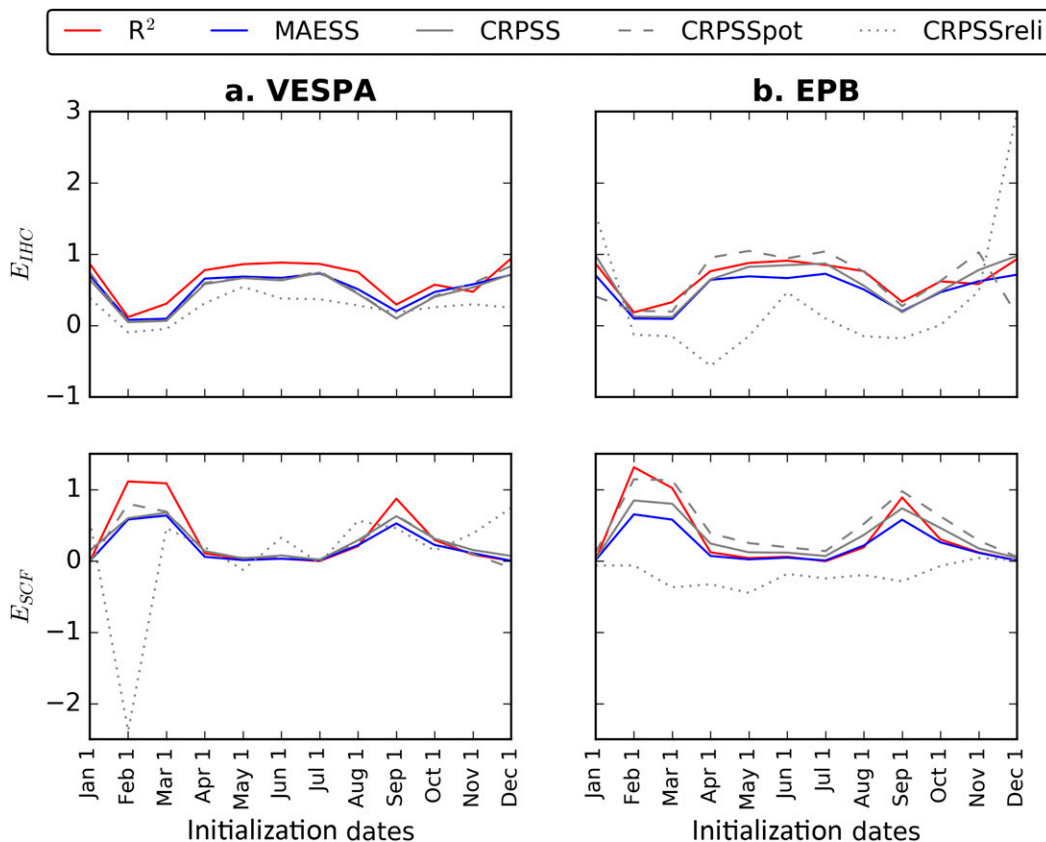
FIG. 6. Streamflow forecast skill elasticities for the (top) IHCs (i.e., E_{IHC}) and (bottom) SCFs (i.e., E_{SCF}) as a function of forecast initialization dates, for hindcasts produced from 1981 to 2010 for the Crystal River (USGS gauge 009081600). These skill elasticities were calculated across a quadrant situated within the 3-month streamflow forecast skill surface plots [from Eqs. (4) and (5)] for several verification scores (R^2 in red, MAESS in blue, CRPSS in gray solid line, CRPSSpot in gray dashed line, and CRPSSreli in gray dotted line). The results are shown for (a) the VESPA approach and (b) the EPB method.

the attribute of the forecast, several scores were computed to evaluate the streamflow forecast quality. The R^2, the mean absolute error skill score (MAESS), and the continuous rank probability skill score (CRPSS) were calculated to evaluate the forecasts' overall performance in terms of the ensemble mean and the entire ensemble. The potential CRPSS (CRPSSpot) was computed to look at the forecasts' resolution and uncertainty, and the CRPSS reliability (CRPSSreli) was computed to look at the forecasts' reliability. The Crystal River (USGS gauge 009081600) will here again be used as a test case to illustrate this part of the results.

Figure 6 presents the IHC and SCF skill elasticities [i.e., E_{IHC} and E_{SCF}; in Fig. 6 (top) and Fig. 6 (bottom), respectively] as a function of forecast initialization date for the Crystal River catchment. These are calculated from Eqs. (4) and (5), for all the mentioned verification scores, for the VESPA approach (Fig. 6a) and the EPB method (Fig. 6b). If we compare the skill elasticities obtained from the VESPA approach with

the skill elasticities obtained from the EPB method, it appears that both methods produce very similar elasticities for the R^2, the MAESS, and the CRPSS. This further confirms the results of the first part of the analysis, which highlighted the similarity of the EPB results to the VESPA results and extends it to multiple attributes of the seasonal streamflow forecasts. However, slight differences between the skill elasticities produced by the two methods can be observed for the CRPSSpot, and significant differences exist for the CRPSSreli. These dissimilarities are discussed further below.

If we now compare the skill elasticities obtained for the various verification scores for both methods, it is clear that the R^2, the MAESS, the CRPSS, and the CRPSSpot give very similar skill elasticities. This hints that those verification scores overall agree on the degree of influence of changes in IHC and SCF skill on the streamflow forecast skill. However, a few dissimilarities can be observed for some of the forecast initialization dates. This is, for example, the case for forecasts made in

the spring and in summer, where the E_{IHC} appears lower for the MAESS and the CRPSS (and the CRPSSpot for the VESPA approach) compared to the E_{IHC} obtained for the R^2 for both methods. It is also apparent for forecasts made on the first of February, March, and September, where the E_{SCF} calculated for the MAESS and the CRPSS (and the CRPSSpot for the VESPA approach) is lower than the E_{SCF} obtained for the R^2 for both methods. For both examples, it infers that improvements in the IHC and the SCF skill could lead to larger improvements in the streamflow forecast skill in terms of the R^2 rather than in terms of the MAESS and the CRPSS (and the CRPSSpot for the VESPA approach). Overall, this indicates that the degree of influence of changes in IHC and SCF skill on the streamflow forecast skill differs relative to the choice of the verification score.

While the R^2, the MAESS, the CRPSS, and the CRPSSpot give a very similar picture, the skill elasticities obtained for the CRPSSreli appear very different, occasionally reaching negative values. These negative values indicate a loss in streamflow forecast skill (in terms of the forecast reliability) as a result of improvements in one of the two predictability sources, while all the other verification scores suggest a gain in streamflow forecast skill (in terms of the forecast ensemble mean and the ensemble overall performance, its resolution, and uncertainty) with improvements in one of the two predictability sources.

The substantial differences in skill elasticities obtained for the CRPSSreli from the VESPA versus EPB method suggest that there are limitations to the ability of EPB to reconstruct the full ensemble information present in VESPA, and of VESPA (applied with relatively small ensembles at the end points) to estimate sensitivities for complex verification scores such as reliability. The reliability verification score is influenced by the combination of bias, spread, and other ensemble properties and exhibits more noisy outcomes here than were obtained for other verification scores. A negative elasticity may occur because the ensemble spread has narrowed without sufficient improvements in bias, for instance. The behavior of the elasticity of reliabilities is even more difficult to diagnose, but we suspect that the presence of noise (erroneous local minima or maxima) or curvature in the associated VESPA skill surface greatly undermines the linear blending techniques.

Overall, these results suggest that improvements in the skill of either of the two predictability sources will impact streamflow forecast skill differently depending on the attribute (i.e., verification score) of the forecast skill that is considered and whether the ensemble mean or the full ensemble is used.

4. Discussion

a. Implications and limitations of the results

W16 introduced the VESPA approach, a sensitivity analysis technique used to pinpoint the dominant predictability source of seasonal streamflow forecasting (i.e., the IHCs and the SCFs), as well as quantifying improvements that can be expected in seasonal streamflow forecast skill as a result of realistic improvements in those key predictability sources. Despite being a powerful sensitivity analysis approach, VESPA presents two key limitations.

1) It is computationally intensive, requiring multiple ensemble hindcasts to define the skill response surface (81 were used in the VESPA paper vs one for the EPB and the EPI techniques).
2) It requires a complex state and forcing blending procedure that may introduce additional uncertainties, biases, or interactions between the predictability sources (Saltelli et al. 2004; Baroni and Tarantola 2014) that are not accounted for or difficult to quantify. This is not necessary in any of the end points required in the two approaches presented here, which rely instead on analyzing a single conventional hindcast dataset that is more likely to be feasible for forecasting centers.

The central aim of this paper was to address the first limitation of the VESPA approach by presenting two computationally inexpensive alternative methods: the EPI and the EPB methods. Both methods successfully identified the dominant predictability source of 3-month streamflow forecasts for a given catchment and forecast initialization date (i.e., given by the orientation of the streamflow forecast skill gradients in the skill surface plots). However, the EPB was more successful in reproducing the VESPA skill elasticities—the exact streamflow forecast skill gradients situated within the skill surface plots (for skill and accuracy verification scores including the R^2, the MAESS, the CRPSS, and the potential CRPSS to a certain extent). These skill elasticities indicate the influence of changes in IHC and SCF skill on streamflow forecast skill.

The new methods, by differing in their setup from the VESPA approach, do not inherit the drawbacks specific to this approach and mentioned above. The EPI and the EPB methods nevertheless have their own limitations.

The EPI (both for the linear and cubic interpolation methods; the latter was not shown) did not fully capture the VESPA skill elasticities because of the nature of the method that produces predefined gradients within the skill surface plots (i.e., defined by the interpolation method used). Additionally, curvature or local minima

or maxima (if any) of the response surface cannot be represented by the EPI method. The EPB, on the other hand, performs better at reflecting curvature in the skill response surface, hence local elasticities between the end points. The EPB method aimed at reproducing VESPA elasticities only by manipulating the output of a single hindcast dataset (interpreted as ESP, reverse ESP, the perfect forecast, and climatology). The EPB method cannot match exactly the forecasts created by the VESPA approach, as it does not account for the idiosyncrasies in model forecast behavior, such as interactions between the predictability sources. Furthermore, it is likely that the more the model investigated is nonlinear or exhibits skill response thresholds, the more the results obtained from the EPB method will differ from the ones obtained from the VESPA approach. These results overall allow that the EPB method can be used as an inexpensive alternative method to the VESPA approach, yet with the potential limitations of the method stated above.

For the first part of the analysis, the streamflow forecast quality was evaluated in terms of the forecasts' skill from the R^2. The use of multiple verification scores is, however, essential to obtain a more complete perspective of forecast quality. Thus, we explored the performance of the two new methods and the VESPA approach for a range of additional verification scores. The results, presented for the EPB method and the VESPA approach, showed differences in the response surfaces obtained for the various verification scores (i.e., the R^2, the MAESS, the CRPSS, and its decomposition). This suggests distinct sensitivities of the seasonal streamflow forecast attributes (i.e., overall performance of the forecast ensemble mean and its full ensemble, forecast resolution, uncertainty, and reliability) to changes in the IHC and SCF skill. Ideally, a sensitivity analysis should be goal oriented, that is, it should be performed with prior knowledge of the intended use of the results (Saltelli et al. 2004; Pappenberger et al. 2010; Baroni and Tarantola 2014), which may favor using one verification score over another.

This paper covered selected limitations of the work presented by W16. However, many areas were left unexplored and could be interesting topics in which to focus future research. First, a major area inherent to model-based sensitivity analyses is that their results are model dependent (Saltelli et al. 2000); thus, the extent to which they can be transferred to reality depends on the model fidelity. The results presented in this paper are specific to the forecasting system and similar systems on which this analysis was based and should be used as an indicator of catchment sensitivities. As noted in W16, an extension of the elasticity analysis to include observations and a model error component would provide valuable insights. Another possible approach could be to use the results from various forecasting systems as input to the sensitivity analysis, in order to achieve a multimodel consensus view of the skill. As shown in Cloke et al. (2017), a multimodel forcing framework can be highly beneficial for streamflow forecasting compared to a single model forecasting approach, provided the models are chosen judiciously so as to provide a rational characterization of forecasting uncertainty. Second, the dependence of blending technique performance versus VESPA on the characteristics of the skill surface (e.g., linear or nonlinear) bears further investigation. Finally, this sensitivity analysis leaves generic the concept of improvements in either of the predictability sources, although the space–time nature of improvements may be consequential. This work could therefore be extended by studying the effect of degradations in the temporal and spatial accuracy of the input data, thereby indicating the relative value of improvements in the spatial or temporal predictability for a specific catchment and a specific time of the year.

b. The wider context

The new strategy of operational forecasting centers is to move toward more integrated operational modeling and forecasting approaches, such as land surface–atmosphere coupled systems, and beyond that, Earth system models. These advances are enabled by the continuous growth of computing capabilities, a better understanding of physical processes and their interactions throughout all compartments of the Earth system, and the availability and use of more and better observation data (i.e., satellite data). Despite all these advances, most forecasts still reflect substantial uncertainty that grows with time and limits the predictability of observed events beyond a few weeks of lead time. The rapid progress has led our systems to be ever more data hungry as increases in model complexity and resolution are sought. These computationally expensive developments are not always feasible; hence, model developers must be creative and constantly weigh the costs and benefits of improving one aspect over another, such as increasing the resolution or complexity of the models (Flato 2011).

In this context, sensitivity analyses appear more than ever as a natural tool to establish priorities in improving predictions based on Earth system modeling. Such analyses are a powerful and valuable tool to support the examination of uncertainty and predictability across spatial and temporal scales and for various applications. They can be used for a large range of activities, including examining model structure,

identifying minimum data standards, establishing priorities for updating forecasting systems, designing field campaigns, and providing realistic insights into the potential benefits of efforts to improve a forecasting system to managers with prior knowledge of their costs (Cloke et al. 2008; Lilburne and Tarantola 2009; W16).

However, sensitivity analyses must be easily reproducible to be effective in supporting each new model or forecast system update, and the results should easily be applied in order to constitute a "continuous learning process" (Baroni and Tarantola 2014). In other words, a sensitivity analysis should be a simple, tractable tool for addressing a multifaceted challenge.

5. Conclusions

This paper presents two computationally inexpensive alternative methods to the VESPA approach for estimating forecast skill sensitivities and elasticities. Of these, the end point blending (EPB) method provides a useful substitute to the VESPA approach. Despite the existence of some differences between the EPB and the VESPA outcomes, the EPB successfully identifies the dominant predictability source (i.e., IHCs and SCFs) of seasonal streamflow forecast skill, for a given catchment and forecast initialization date. The EPB method can additionally reproduce the VESPA forecast skill elasticities, indicating the degree of influence of changes in IHC and SCF skill on the streamflow forecast skill. The paper also draws attention to how the choice of verification score impacts the forecast's sensitivity to improvements made to the predictability sources. With a good understanding of the limitations of the methods, such a sensitivity analysis approach can be a valuable tool to guide future forecasting and modeling developments.

REFERENCES

Baroni, G., and S. Tarantola, 2014: A general probabilistic approach for uncertainty and global sensitivity analysis of deterministic models: A hydrological case study. *Environ. Modell. Software*, **51**, 26–34, doi:10.1016/j.envsoft.2013.09.022.

Bierkens, M. F. P., and L. P. H. van Beek, 2009: Seasonal predictability of European discharge: NAO and hydrological response time. *J. Hydrometeor.*, **10**, 953–968, doi:10.1175/2009JHM1034.1.

Cherry, J., H. Cullen, M. Visbeck, A. Small, and C. Uvo, 2005: Impacts of the North Atlantic Oscillation on Scandinavian hydropower production and energy markets. *Water Resour. Manage.*, **19**, 673–691, doi:10.1007/s11269-005-3279-z.

Chiew, F. H. S., S. L. Zhou, and T. A. McMahon, 2003: Use of seasonal streamflow forecasts in water resources management. *J. Hydrol.*, **270**, 135–144, doi:10.1016/S0022-1694(02)00292-5.

Clark, M. P., M. C. Serreze, and G. J. McCabe, 2001: Historical effects of El Niño and La Niña events on the seasonal evolution of the montane snowpack in the Columbia and Colorado River basins. *Water Resour. Res.*, **37**, 741–757, doi:10.1029/2000WR900305.

Cloke, H. L., F. Pappenberger, and J.-P. Renaud, 2008: Multi-method global sensitivity analysis (MMGSA) for modelling floodplain hydrological processes. *Hydrol. Processes*, **22**, 1660–1674, doi:10.1002/hyp.6734.

——, ——, P. Smith, and F. Wetterhall, 2017: How do I know if I've improved my continental scale flood early warning system? *Environ. Res. Lett.*, **12**, 044006, doi:10.1088/1748-9326/aa625a.

Day, G. N., 1985: Extended streamflow forecasting using NWSRFS. *J. Water Resour. Plann. Manage.*, **111**, 157–170, doi:10.1061/(ASCE)0733-9496(1985)111:2(157).

Demargne, J., and Coauthors, 2014: The science of NOAA's operational Hydrologic Ensemble Forecast Service. *Bull. Amer. Meteor. Soc.*, **95**, 79–98, doi:10.1175/BAMS-D-12-00081.1.

Flato, G. M., 2011: Earth system models: An overview. *Wiley Interdiscip. Rev.: Climate Change*, **2**, 783–800, doi:10.1002/wcc.148.

Hamlet, A. F., D. Huppert, and D. P. Lettenmaier, 2002: Economic value of long-lead streamflow forecasts for Columbia River hydropower. *J. Water Resour. Plann. Manage.*, **128**, 91–101, doi:10.1061/(ASCE)0733-9496(2002)128:2(91).

Kwon, H.-H., C. Brown, K. Xu, and U. Lall, 2009: Seasonal and annual maximum streamflow forecasting using climate information: Application to the Three Gorges Dam in the Yangtze River basin, China. *Hydrol. Sci. J.*, **54**, 582–595, doi:10.1623/hysj.54.3.582.

Li, H., L. Luo, E. F. Wood, and J. Schaake, 2009: The role of initial conditions and forcing uncertainties in seasonal hydrologic forecasting. *J. Geophys. Res.*, **114**, D04114, doi:10.1029/2008JD010969.

Lilburne, L., and S. Tarantola, 2009: Sensitivity analysis of spatial models. *Int. J. Geogr. Inf. Sci.*, **23**, 151–168, doi:10.1080/13658810802094995.

Lins, H. F., 2012: USGS Hydro-Climatic Data Network 2009 (HCDN-2009). USGS Fact Sheet 2012-3047, 4 pp. [Available online at http://pubs.usgs.gov/fs/2012/3047/.]

Luo, L., and E. F. Wood, 2007: Monitoring and predicting the 2007 U.S. drought. *Geophys. Res. Lett.*, **34**, L22702, doi:10.1029/2007GL031673.

MacLeod, D., H. Cloke, F. Pappenberger, and A. Weisheimer, 2016: Evaluating uncertainty in estimates of soil moisture memory with a reverse ensemble approach. *Hydrol. Earth Syst. Sci.*, **20**, 2737–2743, doi:10.5194/hess-20-2737-2016.

Mendoza, P. A., A. W. Wood, E. A. Clark, E. Rothwell, M. P. Clark, B. Nijssen, L. D. Brekke, and J. R. Arnold, 2017: An intercomparison of approaches for improving predictability in operational seasonal streamflow forecasting. *Hydrol. Earth Syst. Sci. Discuss.*, doi:10.5194/hess-2017-60.

Pagano, T., D. Garen, and S. Sorooshian, 2004: Evaluation of official western U.S. seasonal water supply outlooks, 1922–2002. *J. Hydrometeor.*, **5**, 896–909, doi:10.1175/1525-7541(2004)005<0896:EOOWUS>2.0.CO;2.

Paiva, R. C. D., W. Collischonn, M. P. Bonnet, and L. G. G. de Gonçalves, 2012: On the sources of hydrological prediction uncertainty in the Amazon. *Hydrol. Earth Syst. Sci.*, **16**, 3127–3137, doi:10.5194/hess-16-3127-2012.

Pappenberger, F., M. Ratto, and V. Vandenberghe, 2010: Review of sensitivity analysis methods. *Modelling Aspects of Water Approach Directive Implementation*, P. A. Vanrolleghem, Ed., IWA Publishing, 191–265.

Regonda, S. K., B. Rajagopalan, M. Clark, and E. Zagona, 2006: A multimodel ensemble forecast approach: Application to spring seasonal flows in the Gunnison River basin. *Water Resour. Res.*, **42**, W09404, doi:10.1029/2005WR004653.

Saltelli, A., S. Tarantola, and F. Campolongo, 2000: Sensitivity analysis as an ingredient of modeling. *Stat. Sci.*, **15**, 377–395, doi:10.1214/ss/1009213004.

——, ——, ——, and M. Ratto, 2004: *Sensitivity Analysis in Practice: A Guide to Assessing Scientific Models.* John Wiley & Sons, 218 pp.

Shukla, S., and D. P. Lettenmaier, 2011: Seasonal hydrologic prediction in the United States: Understanding the role of initial hydrologic conditions and seasonal climate forecast skill. *Hydrol. Earth Syst. Sci.*, **15**, 3529–3538, doi:10.5194/hess-15-3529-2011.

——, J. Sheffield, E. F. Wood, and D. P. Lettenmaier, 2013: On the sources of global land surface hydrologic predictability. *Hydrol. Earth Syst. Sci.*, **17**, 2781–2796, doi:10.5194/hess-17-2781-2013.

Singla, S., J. P. Céron, E. Martin, F. Regimbeau, M. Déqué, F. Habets, and J. P. Vidal, 2012: Predictability of soil moisture and river flows over France for the spring season. *Hydrol. Earth Syst. Sci.*, **16**, 201–216, doi:10.5194/hess-16-201-2012.

Slater, L. J., G. Villarini, and A. A. Bradley, 2017: Evaluation of the skill of North-American Multi-Model Ensemble (NMME) global climate models in predicting average and extreme precipitation and temperature over the continental USA. *Climate Dyn.*, doi:10.1007/s00382-016-3286-1, in press.

Staudinger, M., and J. Seibert, 2014: Predictability of low flow—An assessment with simulation experiments. *J. Hydrol.*, **519**, 1383–1393, doi:10.1016/j.jhydrol.2014.08.061.

van Dijk, A. I. J. M., J. L. Peña-Arancibia, E. F. Wood, J. Sheffield, and H. E. Beck, 2013: Global analysis of seasonal streamflow predictability using an ensemble prediction system and observations from 6192 small catchments worldwide. *Water Resour. Res.*, **49**, 2729–2746, doi:10.1002/wrcr.20251.

Viel, C., A.-L. Beaulant, J.-M. Soubeyroux, and J.-P. Céron, 2016: How seasonal forecast could help a decision maker: An example of climate service for water resource management. *Adv. Sci. Res.*, **13**, 51–55, doi:10.5194/asr-13-51-2016.

Welles, E., S. Sorooshian, G. Carter, and B. Olsen, 2007: Hydrologic verification: A call for action and collaboration. *Bull. Amer. Meteor. Soc.*, **88**, 503–511, doi:10.1175/BAMS-88-4-503.

Wood, A. W., and D. P. Lettenmaier, 2006: A test bed for new seasonal hydrologic forecasting approaches in the western United States. *Bull. Amer. Meteor. Soc.*, **87**, 1699–1712, doi:10.1175/BAMS-87-12-1699.

——, and ——, 2008: An ensemble approach for attribution of hydrologic prediction uncertainty. *Geophys. Res. Lett.*, **35**, L14401, doi:10.1029/2008GL034648.

——, E. P. Maurer, A. Kumar, and D. P. Lettenmaier, 2002: Long-range experimental hydrologic forecasting for the eastern United States. *J. Geophys. Res.*, **107**, 4429, doi:10.1029/2001JD000659.

——, A. Kumar, and D. P. Lettenmaier, 2005: A retrospective assessment of National Centers for Environmental Prediction climate model–based ensemble hydrologic forecasting in the western United States. *J. Geophys. Res.*, **110**, D04105, doi:10.1029/2004JD004508.

——, T. Hopson, A. Newman, L. Brekke, J. Arnold, and M. Clark, 2016a: Quantifying streamflow forecast skill elasticity to initial condition and climate prediction skill. *J. Hydrometeor.*, **17**, 651–668, doi:10.1175/JHM-D-14-0213.1.

——, T. Pagano, and M. Roos, 2016b: Tracing the origins of ESP. HEPEX, accessed 24 October 2016. [Available online at https://hepex.irstea.fr/tracing-the-origins-of-esp/.]

Yossef, N. C., H. Winsemius, A. Weerts, R. van Beek, and M. F. P. Bierkens, 2013: Skill of a global seasonal streamflow forecasting system, relative roles of initial conditions and meteorological forcing. *Water Resour. Res.*, **49**, 4687–4699, doi:10.1002/wrcr.20350.

Yuan, X., 2016: An experimental seasonal hydrological forecasting system over the Yellow River basin—Part 2: The added value from climate forecast models. *Hydrol. Earth Syst. Sci.*, **20**, 2453–2466, doi:10.5194/hess-20-2453-2016.

——, E. F. Wood, L. Luo, and M. Pan, 2011: A first look at Climate Forecast System version 2 (CFSv2) for hydrological seasonal prediction. *Geophys. Res. Lett.*, **38**, L13402, doi:10.1029/2011GL047792.

——, ——, J. K. Roundy, and M. Pan, 2013: CFSv2-based seasonal hydroclimatic forecasts over the conterminous United States. *J. Climate*, **26**, 4828–4847, doi:10.1175/JCLI-D-12-00683.1.

——, ——, and Z. Ma, 2015: A review on climate-model-based seasonal hydrologic forecasting: Physical understanding and system development. *Wiley Interdiscip. Rev.: Water*, **2**, 523–536, doi:10.1002/wat2.1088.

——, F. Ma, L. Wang, Z. Zheng, Z. Ma, A. Ye, and S. Peng, 2016: An experimental seasonal hydrological forecasting system over the Yellow River basin—Part 1: Understanding the role of initial hydrological conditions. *Hydrol. Earth Syst. Sci.*, **20**, 2437–2451, doi:10.5194/hess-20-2437-2016.