An efficient approach for estimating streamflow forecast skill elasticity

An Efficient Approach for Estimating Streamflow Forecast Skill Elasticity

Louise Arnal[1,2], Andrew W. Wood[3], Elisabeth Stephens[1], Hannah L. Cloke[1,4], Florian Pappenberger[2,5]

[1]Department of Geography and Environmental Science, University of Reading, Reading, UK

[2]ECMWF, European Centre for Medium-Range Weather Forecasts, Shinfield Park, Reading, UK

[3]Research Applications Laboratory, NCAR, Boulder, Colorado

[4]Department of Meteorology, University of Reading, Reading, UK

[5]School of Geographical Sciences, University of Bristol, Bristol, UK

Correspondence to:

Louise Arnal

Department of Geography and Environmental Science

School of Archaeology, Geography and Environmental Science

The University of Reading

Whiteknights, PO Box 227

Reading

RG6 6AB

UK

l.l.s.arnal@pgr.reading.ac.uk; louise.arnal@ecmwf.int

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 modelling 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 modelling systems. This has led to a shift in the strategy of operational forecasting centres towards more integrated modelling and forecasting approaches, such as coupled systems and Earth System Models (ESMs), with the final aim to extend the limits of predictability (i.e., sub-seasonal 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. 2016; 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,

71 snowpack and the current streamflow). IHCs are one of the most important predictability
72 sources of seasonal streamflow forecasts and were thus the starting point for the
73 development of the Ensemble Streamflow Predictions (ESP) approach in the 1970s (Wood et
74 al. 2016b). The ESP was first developed and used for reservoir management purposes. It is
75 produced by running a hydrological model with observed meteorological inputs to produce
76 current observed IHCs, from which the forecast is started, and the forcing over the forecast
77 period is done with an ensemble of historical meteorological observations (Day 1985). The
78 ESP method assumes that the model states to initialise a forecast are perfectly estimated,
79 while the future climate is completely unknown. However, the skill of the ESP decreases
80 significantly after one to a few months of lead time over most parts of the world due to a
81 decrease in the land surface memory with time. The achievable predictability from the ESP
82 thus depends on the persistence of the IHCs, which can vary as a function of the season (i.e.,
83 the transition between dry and wet seasons will for example be hard to forecast) and the
84 location and size of the catchment (i.e., the streamflow in a large catchment with a slow
85 response time and/or situated in a region with negligible precipitation inputs during the
86 forecast period will for example be easier to forecast; Wood and Lettenmaier 2008; Shukla et
87 al. 2013; van Dijk et al. 2013; Yuan et al. 2015).

88 More recently seasonal climate predictability derived from large scale climate precursors
89 (e.g., the 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 initialising a hydrological model to estimate IHCs and forcing the model with inputs based on seasonal climate forecasts (SCF; of temperature and precipitation) instead of historical observations. Their skill is also still limited, due to 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 that is being tackled by improving IHCs and the SCFs using a variety of techniques. Techniques include model developments and data assimilation and can be more or less expensive. 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 IHCs and SCFs 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 initialise a prediction forced by a single set of observed meteorological inputs. Typically, the input uncertainty damps out 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), 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 skilful 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 IHCs and SCFs 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', i.e., 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 initialisation month of a forecast, the response surface was defined through the use of dozens of multi-decadal 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 also was 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., R^2) for the analysis. An ensemble forecast has many attributes: e.g., the skill, the reliability, the resolution and the uncertainty of the forecast, among others. In order to obtain a complete picture of the forecast quality,

146 the scores should encompass many of these attributes as each verification score will give us
147 different information about the forecast quality.

148 The drawbacks of VESPA motivate us to assess two computationally inexpensive methods
149 of estimating the forecast skill elasticities, using only the original ESP and reverse-ESP results
150 that depend on the single hindcast series as mentioned above. The two methods are termed
151 End Point Interpolation (EPI) and End Point Blending (EPB). In the first part of this paper, we
152 compare results from the two methods tested on 18 catchments of the CONUS to the original
153 results from the VESPA, using a single verification score. The objective of this part is to
154 investigate whether the new methods can discriminate the influence of IHCs and SCFs errors
155 on seasonal streamflow forecasting uncertainties and to assess the ability of those new
156 methods to correctly estimate the forecast skill elasticities. In the second part, additional
157 verification scores are applied for streamflow forecast verification, supporting the second
158 objective of the paper, which is to explore the sensitivity of the results obtained from the two
159 new methods and the VESPA approach to the choice of the verification score.

160 **2. Methods, data and evaluation strategy**

161 **a. The VESPA approach**

162 In this work, as in W16, the term 'perfect' refers to current observed meteorological
163 data and the term climatological refers to the whole distribution of historical observed data.
164 Figure 1 presents the ESP (Figure 1a), the reverse-ESP (Figure 1b), the climatology (Figure 1c)

and the VESPA forecast (Figure 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.

The VESPA aims to produce streamflow forecasts from IHCs and SCFs with an uncertainty situated between the 'perfect' and the climatological uncertainty (Figure 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' 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 (Figure 2), interpolating forecast skill results from different IHCs and SCFs 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):

$$SCF\ skill = 100 \times \left(1 - w_{SCF}{}^{2}\right) \quad (1)$$

$$IHC\ skill = 100 \times \left(1 - w_{IHC}{}^{2}\right) \quad (2)$$

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, which skill can be plotted on the skill surface plot (black crosses in Figure 2). The blue circles are the end points of the VESPA forecasts: the reverse-ESP (revESP in Figure 2), the 'perfect' forecasts, the ESP and the climatology (climo in Figure 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 calibrated and validated against observed daily streamflow from the US Geological Survey (USGS). Eighty-one skill variations of a 30 year hindcast (from 1981 to 2010) were produced for 424 catchments in the CONUS, starting at the beginning of each month (i.e., forecast initialisation dates), with lead times up to six months.

b. Alternative methods to the VESPA approach

In this paper we present two alternative methods of the VESPA approach, called the End Point Interpolation (EPI) and the End Point Blending (EPB). These methods aim to reproduce the response surface obtained from the VESPA approach, by using the same 30 year hindcast ensembles produced by W16, aggregated over the first three months with zero lead time for each initialisation 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).

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

The EPI produces a response surface by interpolating the forecast skill of the end points throughout the skill 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 initialisation date and for each catchment.

2) End point Blending (EPB)

The EPB generates hindcasts for each w_{SCF} - w_{IHC} combination (i.e., each cross in Figure 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:

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

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 percent 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 initialisation date and for each catchment.

In order 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., Eq. (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 IHCs and SCFs 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 IHCs and SCFs 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 Pearson product moment correlation coefficient (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 centre of the response surface, according to the following equations:

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

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

The numerators, expressed as S(F[-])-S(F[-]), 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 Figure 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 initialisation date.

The only difference between Eq. (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 than in W16, as 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.

290 **2) What is the sensitivity of the response surface to the choice of the**

291 **verification score?**

292 In order to investigate the second objective of this paper, several verification scores were

293 calculated for each method (i.e., the EPI, the EPB and the VESPA approach). These scores were

294 selected in order to cover key attributes of the forecasts verified, they include:

295 • the Mean Absolute Error (MAE) of forecast ensemble means, relative to the 'perfect'

296 forecasts,

297 • the Continuous Rank Probability Score (CRPS) and its decomposition:

298 ○ the potential CRPS (CRPSpot): where CRPSpot = resolution - uncertainty,

299 ○ the reliability part of the CRPS (CRPSreli).

300 The potential CRPS is the CRPS value that a forecast with perfect reliability would have. The

301 uncertainty is the variability of the observations and the resolution is the ability of the forecast

302 to distinguish situations with distinctly different frequencies of occurrence. The CRPS

303 reliability is a measure of the bias and the spread of the system.

304 The CRPS was chosen as it is a widely used score to assess the overall quality of an

305 ensemble hydrometeorological forecast. The CRPS moreover has the advantage that it can be

306 decomposed in different scores to look at many attributes of an ensemble forecast. The CRPS

307 score 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:

$$Skill\ score_{forecast} = 1 - \frac{score_{forecast}}{score_{reference}} \quad (6)$$

Where the $score_{reference}$ is the score of the climatology point, for each method, each initialisation date and each catchment. A perfect forecast results in a forecast skill score of one and a forecast with equal quality as the reference forecast corresponds to a skill score of zero. Any forecasts with less quality than the reference forecast produces 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 Eq. (4) and (5), for all three methods and for the 3-month streamflow forecasts produced for each catchment and forecast initialisation 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).

2. Results

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

For the first part of this study, the Crystal River (CO; 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 (Figure 4a), the linear EPI (Figure 4b) and the EPB methods (Figure 4c), from R^2 for the 3-month streamflow forecast for the Crystal River, for initialisations on the first of each month (each row on Figure 4). Figures 4d and 4e show the differences between the skill surface plots obtained for the VESPA and the EPI methods, and the VESPA and the EPB methods, respectively. A first visual comparison of the skill surface plots obtained from the linear EPI method (Figure 4b) and the EPB method (Figure 4c) with those obtained from the VESPA approach (Figure 4a) for the Crystal River tells us that the skill surface plots obtained from all three methods are very similar. For each initialisation date, the orientation of the gradients in streamflow forecast skill appears identical. The EPI and the EPB methods seem to correctly indicate the dominant

343 predictability source on the 3-month streamflow forecast skill, for each initialisation date for
344 this catchment. Similar results were obtained for the other 17 catchments (see
345 Supplementary Figures 1 to 17). Forecasts made on the 1st of February, March and September
346 show a sensitivity to the SCF skill (i.e., horizontal or near to horizontal orientation of the
347 streamflow forecast skill gradients), while all other forecasts are dominantly sensitive to the
348 IHC skill (i.e., vertical or near to vertical orientation of the streamflow forecast skill gradients).

349 The gradients in streamflow forecast skill contained in the EPI skill surface plots (Figure
350 4b) differ moderately from the gradients obtained from the VESPA approach (Figure 4a). This
351 can be observed in Figure 4d, showing the differences between the skill surface plots obtained
352 for both methods. The VESPA approach gives very strong gradients causing a rapid decrease
353 in streamflow forecast skill with a decrease in one of the predictability sources' skill,
354 depending on the initialisation date. In comparison, the EPI method indicates a gradual
355 decrease in streamflow forecast skill with a decrease in one of the two predictability sources,
356 depending on the initialisation date. The streamflow forecast skill gradients produced by the
357 EPI method are a reflection of the interpolation method used (i.e., here linear), and because
358 the corner points lack information about describing curvature of the surface at interior points,
359 they cannot fully capture non-linearities in the skill gradients across the skill surface. For some
360 interior points, this limitation of the EPI method could estimate very different skill elasticities
361 than those obtained from the VESPA approach.

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

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

379 The value of the SCF skill elasticity (i.e., E_{SCF}) in relation to the value of the IHC skill
380 elasticity (i.e., E_{IHC}), for a given method, indicates the dominant predictability source on the

381 3-month streamflow forecast skill (here calculated from the R^2). For a selected method, equal
382 SCF and IHC skill elasticity values signifies that equal improvements in both the SCFs and the
383 IHCs will lead to equal improvements in the streamflow forecast skill. If E_{SCF} is superior
384 [inferior] to E_{IHC}, it reflects a larger potential increase in streamflow forecast skill by improving
385 the SCFs [IHCs]. Although the EPI method almost always indicates the same dominant
386 predictability source as the two other methods, the degree of influence of changes in IHC and
387 SCF skill on the streamflow forecast skill (i.e., the exact values of the skill elasticities) often
388 differs. For many catchments and forecast initialisation dates, the EPI appears to
389 underestimate the skill elasticities produced by the VESPA method.

390 The nine different catchments for which the skill elasticities are presented in Figure 5
391 display three different types of behaviours, best captured by the VESPA approach and the EPB
392 method. For the three catchments on the leftmost column of Figure 5, improvements in the
393 IHCs would yield the highest improvements in the 3-month streamflow forecast skill for spring
394 to summer initialisations (April-August for the Crystal River, March-July for the Fish River and
395 March-June for the Middle Branch Escanaba River) and in the winter (October-January for the
396 Crystal River, November-December for the Fish River and in December for the Middle Branch
397 Escanaba River). SCF improvements would lead to better 3-month streamflow forecast skill
398 for forecasts initialised in the late winter and summer to fall (February-March and September
399 for the Crystal River, February and August-October for the Fish River and January-February

400 and July-September for the Middle Branch Escanaba River). For the three catchments in the
401 middle column of Figure 5, a notable feature is that the 3-month streamflow forecast skill
402 would benefit from SCF improvements for summer initialisations (June-September for the
403 Chattooga and the Nantahala Rivers and July-September for the New River). Finally, for the
404 three catchments of the rightmost column of Figure 5, the 3-month streamflow forecast skill
405 would benefit from improvements in the SCFs for all initialisation dates. This is true with the
406 exception of forecasts initialised in December for East Fork Shoal Creek. It is important to note
407 that there is uncertainty around these estimates. However, this is a good first indication of
408 the sensitivity of 3-month streamflow forecast skill (measured from the R^2) to IHCs and SCFs
409 errors, for each forecast initialisation date and each catchment.

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

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

In order to investigate the sensitivity of the response surface to the choice of the verification score, and therefore to the attribute of the forecast, several scores were computed to evaluate the streamflow forecast quality. The R^2, the MAE skill score (MAESS) and the 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) to look at the forecasts' reliability. Crystal River (CO; 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 the top two plots and the bottom two plots of Figure 6, respectively) as a function of forecast initialisation date, for the Crystal River catchment. These are calculated from Eq. (4) and (5), for all the mentioned verification scores, for the VESPA approach (Figure 6a, the two leftmost plots) and the EPB method (Figure 6b, the two rightmost plots). 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 initialisation 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 1st 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 than in terms of the MAESS and the CRPSS (and the CRPSSpot for the VESPA approach). This overall 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 behaviour 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.

3. Discussion

a. Implications and limitations of the results

W16 introduced the variational ensemble streamflow prediction assessment (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 versus 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 analysing a single conventional hindcast dataset that is more likely to be feasible for forecasting centres.

The central aim of this paper was to address the first limitation of the VESPA approach by presenting two computationally inexpensive alternative methods: the End Point Interpolation (EPI) and the End Point Blending (EPB) methods. Both methods successfully identified the dominant predictability source of 3-month streamflow forecasts for a given catchment and forecast initialisation 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, due to the nature of the method which 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 behaviour, such as interactions between the predictability sources. Furthermore, it is likely that the more the model investigated is non-linear 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

530 seasonal streamflow forecast attributes (i.e., overall performance of the forecast ensemble
531 mean and its full ensemble, forecast resolution, uncertainty and reliability) to changes in the
532 IHC and SCF skill. Ideally, a sensitivity analysis should be goal-oriented – i.e., it should be
533 performed with prior knowledge of the intended use of the results (Saltelli et al. 2004;
534 Pappenberger et al. 2010; Baroni and Tarantola 2014), which may favour using one
535 verification score over another.

536 This paper covered selected limitations of the work presented by W16. Many areas were
537 however left unexplored and could be interesting topics to focus future research. Firstly, a
538 major area inherent to model-based sensitivity analyses is that their results are model
539 dependent (Saltelli et al. 2000), thus the extent to which they can be transferred to reality
540 depends on the model fidelity. The results presented in this paper are specific to the
541 forecasting system, and similar systems, on which this analysis was based and should be used
542 as an indicator of catchment sensitivities. As noted in W16, an extension of the elasticity
543 analysis to include observations and a model error component would provide valuable
544 insights. Another possible approach could be to use the results from various forecasting
545 systems as input to the sensitivity analysis, in order to achieve a multi model consensus view
546 of the skill. As shown in Cloke et al. (2017), a multi model forcing framework can be highly
547 beneficial for streamflow forecasting compared to a single model forecasting approach,
548 provided the models are chosen judiciously so as to provide a rational characterisation of

549 forecasting uncertainty. Secondly, the dependence of blending technique performance versus
550 VESPA on the characteristics of the skill surface (e.g., linear or non-linear) bears further
551 investigation. Finally, this sensitivity analysis leaves generic the concept of improvements in
552 either of the predictability sources, although the space-time nature of improvements may be
553 consequential. This work could therefore be extended by studying the effect of degradations
554 in the temporal and spatial accuracy of the input data, thereby indicating the relative value
555 of improvements in the spatial or temporal predictability for a specific catchment and a
556 specific time of the year.

557 **b. The wider context**

558 The new strategy of operational forecasting centres is to move towards more
559 integrated operational modelling and forecasting approaches, such as land surface-
560 atmosphere coupled systems, and beyond that, Earth System Models. These advances are
561 enabled by the continuous growth of computing capabilities, a better understanding of
562 physical processes and their interactions throughout all compartments of the Earth, and the
563 availability and use of more and better observation data (i.e., satellite data). Despite all these
564 advances, most forecasts still reflect substantial uncertainty that grows with time and limits
565 the predictability of observed events beyond a few weeks of lead time. The rapid progress
566 has led our systems to be ever more data hungry as increases in model complexity and
567 resolution are sought. These computationally expensive developments are not always

568 feasible, hence, model developers must be creative and constantly weigh the costs and
569 benefits of improving one aspect over another, such as increasing the resolution or
570 complexity of the models (Flato 2011).

571 In this context, sensitivity analyses appear more than ever as a natural tool to establish
572 priorities in improving predictions based on Earth System Modelling. Such analyses are a
573 powerful and valuable tool to support the examination of uncertainty and predictability
574 across spatial and temporal scales and for various applications. They can be used for a large
575 range of activities, including: examining model structure, identifying minimum data
576 standards, establishing priorities for updating forecasting systems, designing field campaigns
577 and providing realistic insights into the potential benefits of efforts to improve a forecasting
578 system to managers with prior knowledge of their costs (Cloke et al. 2008; Lilburne and
579 Tarantola 2009; W16).

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

4. 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., the initial hydrological conditions [IHCs] and the seasonal climate forecasts [SCFs]) of seasonal streamflow forecast skill, for a given catchment and forecast initialisation 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 represent a valuable tool to guide future forecasting and modelling developments.

Acknowledgments. The authors gratefully acknowledge financial support from the Horizon 2020 IMPREX project (grant agreement 641811) (project IMPREX: www.imprex.eu). E. Stephens' time was funded by the Leverhulme Early Career Fellowship ECF-2013-492. We also acknowledge high-performance computing support from Yellowstone (ark:/85065/d7wd3xhc) provided by NCAR's Computational and Information Systems Laboratory, sponsored by the National Science Foundation, and financial support from NCAR's Visiting Scientist Program. Lastly, we are thankful for support from the US Bureau of Reclamation under Cooperative Agreement R11AC80816 and from the US Army Corps of Engineers (USACE) Climate Preparedness and Resilience Program.

5. References

Baroni, G., and S. Tarantola, 2014: A General Probabilistic Approach for uncertainty and global sensitivity analysis of deterministic models: A hydrological case study. *Environmental Modelling & 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,** 4, 953–68, doi:10.1175/2009JHM1034.1.

Cherry, J., H. Cullen, M. Visbeck, A. Small, and C. Uvo, 2016: Impacts of the North Atlantic Oscillation on Scandinavian Hydropower Production and Energy Markets. *Water Resources Management*, **19**, 6, 673–91, 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. *Journal of Hydrology*, **270**, 1–2, 135–44, doi:10.1016/S0022-1694(02)00292-5.

Clark, M. P., M. C. Serreze, and G. J. McCabe, 2001: Historical Effects of El Nino and La Nina Events on the Seasonal Evolution of the Montane Snowpack in the Columbia and Colorado River Basins. *Water Resources Research*, **37**, 3, 741–57, doi:10.1029/2000WR900305.

635 Cloke, H. L., F. Pappenberger, and J-P. Renaud, 2008: Multi-method global sensitivity
636 analysis (MMGSA) for modelling floodplain hydrological processes. *Hydrological processes*,
637 **22**, 11, 1660-1674, doi:10.1002/hyp.6734.

638 Cloke H. L., F. Pappenberger, P. Smith, and F. Wetterhall, 2017: How do I know if I've
639 improved my continental scale flood early warning system? *Environmental Research Letters*,
640 (accepted).

641 Day, G. N., 1985: Extended Streamflow Forecasting Using NWSRFS. *Journal of Water*
642 *Resources Planning and Management*, **111**, 2, 157–170, doi:10.1061/(ASCE)0733-
643 9496(1985)111:2(157).

644 Demargne, J., and Coauthors, 2014: The Science of NOAA's Operational Hydrologic
645 Ensemble Forecast Service. *Bull. Amer. Meteor. Soc.*, **95**, 1, 79–98, doi:10.1175/BAMS-D-12-
646 00081.1.

647 Flato, G. M., 2011: Earth System Models: An Overview. *Wiley Interdisciplinary Reviews:*
648 *Climate Change*, **2**, 6, 783–800, doi:10.1002/wcc.148.

649 Hamlet, A. F., D. Huppert, and D. P. Lettenmaier, 2002: Economic Value of Long-Lead
650 Streamflow Forecasts for Columbia River Hydropower. *Journal of Water Resources Planning*
651 *and Management*, **128**, 2, 91–101, doi:10.1061/(ASCE)0733-9496(2002)128:2(91).

652 Kwon, H-H., C. Brown, K. Xu, and U. Lall, 2009: Seasonal and Annual Maximum Streamflow
653 Forecasting Using Climate Information: Application to the Three Gorges Dam in the Yangtze
654 River Basin, China / Prévision D'écoulements Saisonnier et Maximum Annuel à L'aide
655 D'informations Climatiques: Application Au Barrage Des Trois Gorges Dans Le Bassin Du
656 Fleuve Yangtze, Chine. *Hydrological Sciences Journal*, **54**, 3, 582–95,
657 doi:10.1623/hysj.54.3.582.

658 Li, H., L. Luo, E. F. Wood, J. Schaake, 2009: The role of initial conditions and forcing
659 uncertainties in seasonal hydrologic forecasting. *Journal of Geophysical Research:*
660 *Atmospheres*, **114**, D04114, doi:10.1029/2008JD010969.

661 Lilburne, L., and S. Tarantola, 2009: Sensitivity analysis of spatial models. *International*
662 *Journal of Geographical Information Science*, **23**, 2, 151-168,
663 doi:10.1080/13658810802094995.

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

667 Luo, L., and E. F. Wood, 2007: Monitoring and Predicting the 2007 U.S. Drought. *Geophysical*
668 *Research Letters*, **34**, L22702, doi:10.1029/2007GL031673.

669 MacLeod, D., H. Cloke, F. Pappenberger, and A. Weisheimer, 2016: Evaluating Uncertainty in
670 Estimates of Soil Moisture Memory with a Reverse Ensemble Approach. *Hydrology and*
671 *Earth System Sciences*, **20**, 7, 2737–43, doi:10.5194/hess-20-2737-2016.

672 Mendoza, P. A., A. W. Wood, E. A. Clark, E. Rothwell, M. P. Clark, B. Nijssen, L. D. Brekke, and
673 J. R. Arnold, 2017: An intercomparison of approaches for improving predictability in
674 operational seasonal streamflow forecasting, *Hydrology and Earth System Sciences*
675 *Discussions* (in review).

676 Pagano, T., D. Garen, and S. Sorooshian, 2004: Evaluation of official western US seasonal
677 water supply outlooks, 1922–2002. *J. Hydrometeor.*, **5**, 5, 896–909, doi: 10.1175/1525-
678 7541(2004)005<0896:EOOWUS>2.0.CO;2.

679 Paiva, R. C. D., W. Collischonn, M. P. Bonnet, and L. G. G. de Gonçalves, 2012: On the
680 sources of hydrological prediction uncertainty in the Amazon. *Hydrology and Earth System*
681 *Sciences*, **16**, 9, 3127-3137, doi:10.5194/hess-16-3127-2012.

682 Pappenberger, F., M. Ratto, and V. Vandenberghe, 2010: Review of sensitivity analysis
683 methods. *Modelling aspects of water approach directive implementation*, P. A.
684 Vanrolleghem, IWA Publishing, 191-265.

685 Regonda, S. K., B. Rajagopalan, M. Clark, and E. Zagona, 2006: A Multimodel Ensemble
686 Forecast Approach: Application to Spring Seasonal Flows in the Gunnison River Basin. *Water*
687 *Resources Research*, **42**, W09404, doi:10.1029/2005WR004653.

688 Saltelli, A., S. Tarantola, and F. Campolongo, 2000: Sensitivity analysis as an ingredient of
689 modeling. *Statistical Science*, **15**, 4, 377-395.

690 Saltelli, A., S. Tarantola, F. Campolongo, and M. Ratto, 2004: *Sensitivity analysis in practice:*
691 *a guide to assessing scientific models*. John Wiley & Sons, 218 pp.

692 Shukla, S., and D. P. Lettenmaier, 2011: Seasonal hydrologic prediction in the United States:
693 understanding the role of initial hydrologic conditions and seasonal climate forecast skill.
694 *Hydrology and Earth System Sciences*, **15**, 11, 3529-3538, doi:10.5194/hess-15-3529-2011.

695 Shukla, S., J. Sheffield, E. F. Wood and D. P. Lettenmaier, 2013: On the sources of global land
696 surface hydrologic predictability. *Hydrology and Earth System Sciences*, **17,** 7, 2781-2796,
697 doi:10.5194/hess-17-2781-2013.

698 Singla, S., J. P. Céron, E. Martin, F. Regimbeau, M. Déqué, F. Habets, J. P. Vidal, 2012:
699 Predictability of soil moisture and river flows over France for the spring season. *Hydrology*
700 *and Earth System Sciences*, **16**, 1, 201-216, doi:10.5194/hess-16-201-2012.

701 Slater, L. J., G. Villarini, and A. A. Bradley, 2016: Evaluation of the Skill of North-American
702 Multi-Model Ensemble (NMME) Global Climate Models in Predicting Average and Extreme

703 Precipitation and Temperature over the Continental USA. *Climate Dynamics*, 1-16,
704 doi:10.1007/s00382-016-3286-1.

705 Staudinger, M., and J. Seibert, 2014: Predictability of low flow–An assessment with
706 simulation experiments. *Journal of Hydrology*, **519**, 1383-1393,
707 doi:10.1016/j.jhydrol.2014.08.061.

708 van Dijk, A. I. J. M., J. L. Peña-Arancibia, E. F. Wood, J. Sheffield, and H. E. Beck, 2013: Global
709 Analysis of Seasonal Streamflow Predictability Using an Ensemble Prediction System and
710 Observations from 6192 Small Catchments Worldwide. *Water Resources Research*, **49**, 5,
711 2729–46, doi:10.1002/wrcr.20251.

712 Viel, C., A-L. Beaulant, J-M. Soubeyroux, and J-P. Céron, 2016: How Seasonal Forecast Could
713 Help a Decision Maker: An Example of Climate Service for Water Resource Management.
714 *Advances in Science and Research*, **13**, 51–55, doi:10.5194/asr-13-51-2016.

715 Welles, E., S. Sorooshian, G. Carter, and B. Olsen, 2007: Hydrologic Verification: A Call for
716 Action and Collaboration. *Bull. Amer. Meteor. Soc.*, **88**, 4, 503–11, doi:10.1175/BAMS-88-4-
717 503.

718 Wood, A. W., E. P. Maurer, A. Kumar, and D. P. Lettenmaier, 2002: Long-Range Experimental
719 Hydrologic Forecasting for the Eastern United States. *Journal of Geophysical Research:*
720 *Atmospheres*, **107**, D20, doi:10.1029/2001JD000659.

721 Wood, A. W., A. Kumar, and D. P. Lettenmaier, 2005: A Retrospective Assessment of
722 National Centers for Environmental Prediction Climate Model–based Ensemble Hydrologic
723 Forecasting in the Western United States. *Journal of Geophysical Research: Atmospheres*,
724 **110**, D04105, doi:10.1029/2004JD004508.

725 Wood, A. W., and D.P. Lettenmaier, 2006: A new approach for seasonal hydrologic
726 forecasting in the western U.S. *Bull. Amer. Meteor. Soc.*, **87**, 12, 1699-1712,
727 doi:10.1175/BAMS-87-12-1699.

728 Wood, A. W., and D. P. Lettenmaier, 2008: An ensemble approach for attribution of
729 hydrologic prediction uncertainty. *Geophysical Research Letters*, **35**, 14,
730 doi:10.1029/2008GL034648.

731 Wood, A. W., T. Hopson, A. Newman, L. Brekke, J. Arnold, and M. Clark, 2016: Quantifying
732 Streamflow Forecast Skill Elasticity to Initial Condition and Climate Prediction Skill. *J.*
733 *Hydrometeor.*, **17**, 2, 651–668, doi:10.1175/JHM-D-14-0213.1.

734 Wood, A. W., T. Pagano, and M. Roos, 2016: Tracing The Origins of ESP. Accessed 24
735 October 2016. [Available online at https://hepex.irstea.fr/tracing-the-origins-of-esp/.]

736 Yossef, N. C., H. Winsemius, A. Weerts, R. van Beek, and M. F. P. Bierkens, 2013: Skill of a
737 global seasonal streamflow forecasting system, relative roles of initial conditions and

meteorological forcing. *Water Resources Research*, **49**, 8, 4687-4699, doi:10.1002/wrcr.20350.

Yuan, X., E. F. Wood, L. Luo, and M. Pan, 2011: A First Look at Climate Forecast System Version 2 (CFSv2) for Hydrological Seasonal Prediction. *Geophysical Research Letters*, **38**, L13402, doi:10.1029/2011GL047792.

Yuan, X., E. F. Wood, J. K. Roundy, and M. Pan, 2013: CFSv2-Based Seasonal Hydroclimatic Forecasts over the Conterminous United States. *J. Climate*, **26**, 13, 4828–47, doi:10.1175/JCLI-D-12-00683.1.

Yuan, X., E. F. Wood, and Z. Ma, 2015: A Review on Climate-Model-Based Seasonal Hydrologic Forecasting: Physical Understanding and System Development. *Wiley Interdisciplinary Reviews: Water*, **2**, 5, 523–36, doi:10.1002/wat2.1088.

Yuan, X., 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. *Hydrology and Earth System Sciences*, **20**, 2437-2451, doi:10.5194/hess-20-2437-2016.

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

Figure 1 Schematic of a. the ESP, b. the reverse-ESP, c. the climatology and d. the VESPA (this figure is adapted from Figure 3 from W16).

Figure 2 Schematic of a skill surface plot. The y and the 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.

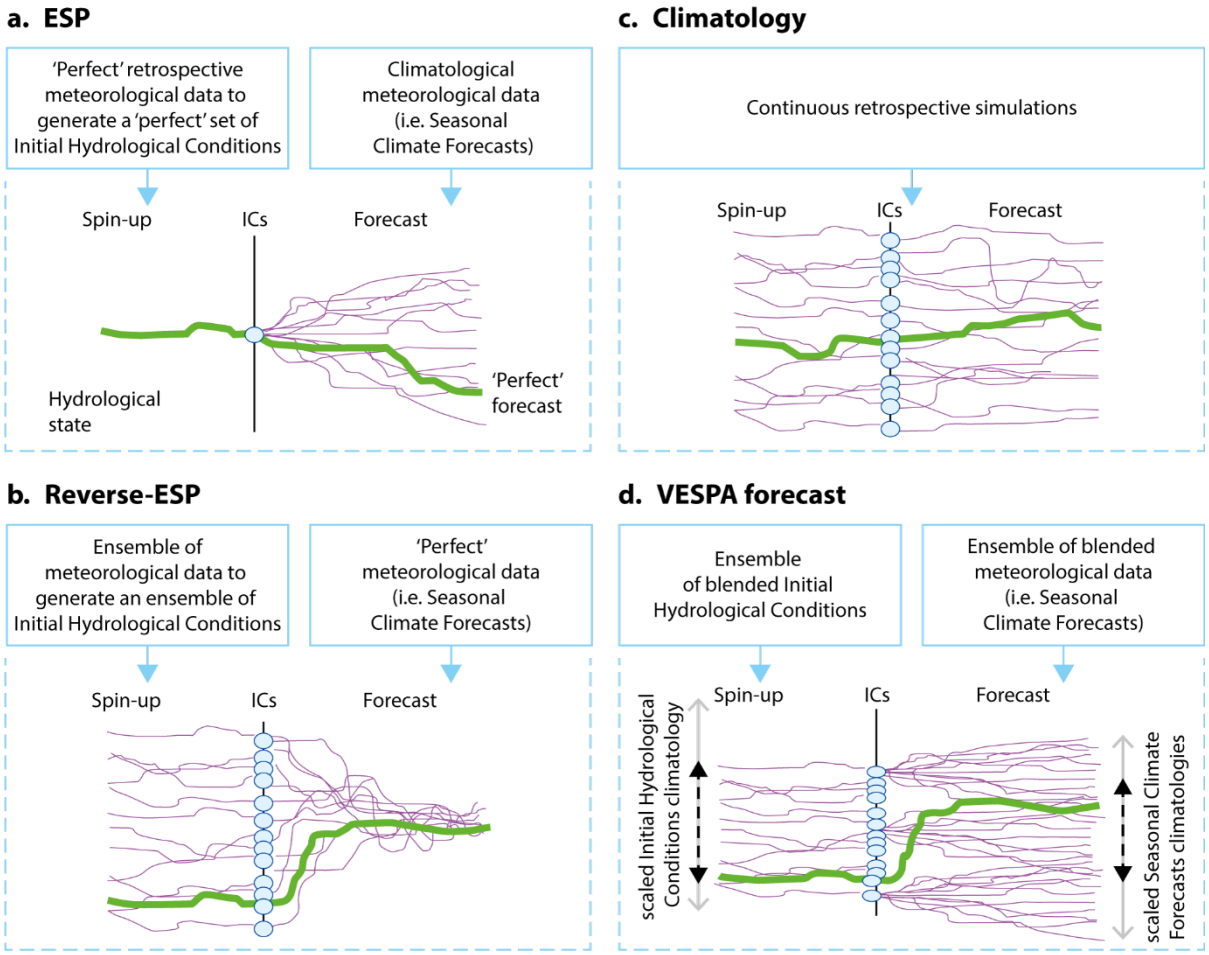
Figure 3 Map of the 18 catchments of the CONUS selected for the analysis, and the HCDN regions (dark blue outlines).

Figure 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 1981-2010 for the Crystal River (CO; USGS gauge 009081600), with forecast initialisations on the first day of each month. Differences between the skill surface plots obtained for the d. VESPA and linear EPI methods and the e. VESPA and EPB methods are also shown.

Figure 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 (in red), the linear EPI method (in grey) and the EPB method (in blue; using Eq. (4) and (5)*). Each plot shows the evolution of the IHC and SCF skill*

774 *elasticities with the initialisation date for a given catchment. The climatological regions of the*
775 *catchments are indicated in the plots' headings. The skill surface plots from which these skill*
776 *elasticities were calculated are presented in Figure 4 and Supplementary Figures 1 to 17.*

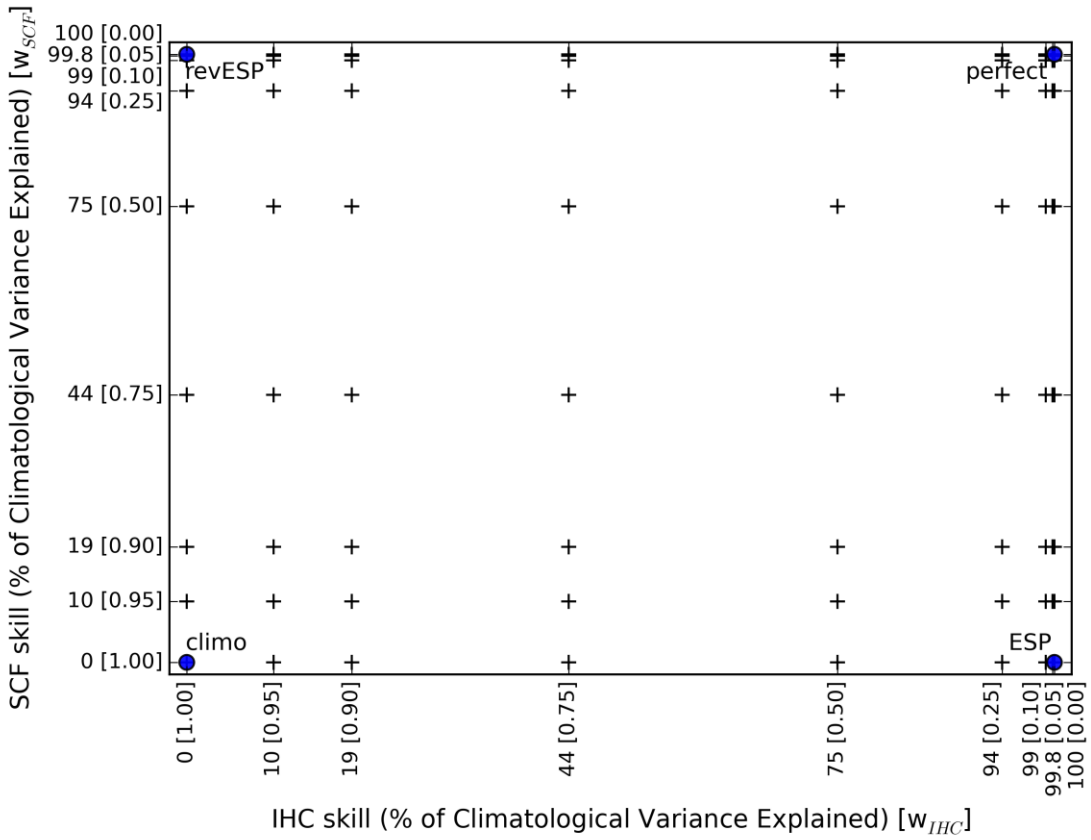
777 *Figure 6 Streamflow forecast skill elasticities for the IHCs (i.e., E_{IHC}, top two plots) and the SCFs*
778 *(i.e., E_{SCF}, bottom two plots) as a function of forecast initialisation dates, for hindcasts*
779 *produced from 1981-2010 for the Crystal River (CO; USGS gauge 009081600). These skill*
780 *elasticities were calculated across a quadrant situated within the 3-month streamflow*
781 *forecast skill surface plots (from Eq.* (4) and (5)*) for several verification scores (the R^2 in red,*
782 *the MAE skill score [MAESS] in blue, the CRPSS in grey solid line, the potential CRPSS*
783 *[CRPSSpot] in grey dashed line and the CRPSS reliability [CRPSSreli] in grey dotted line). The*
784 *results are shown for a. the VESPA approach (two leftmost plots) and b. the EPB method (two*
785 *rightmost plots).*



786

787 *Figure 1 Schematic of a. the ESP, b. the reverse-ESP, c. the climatology and d. the VESPA (this*

788 *figure is adapted from Figure 3 from W16).*



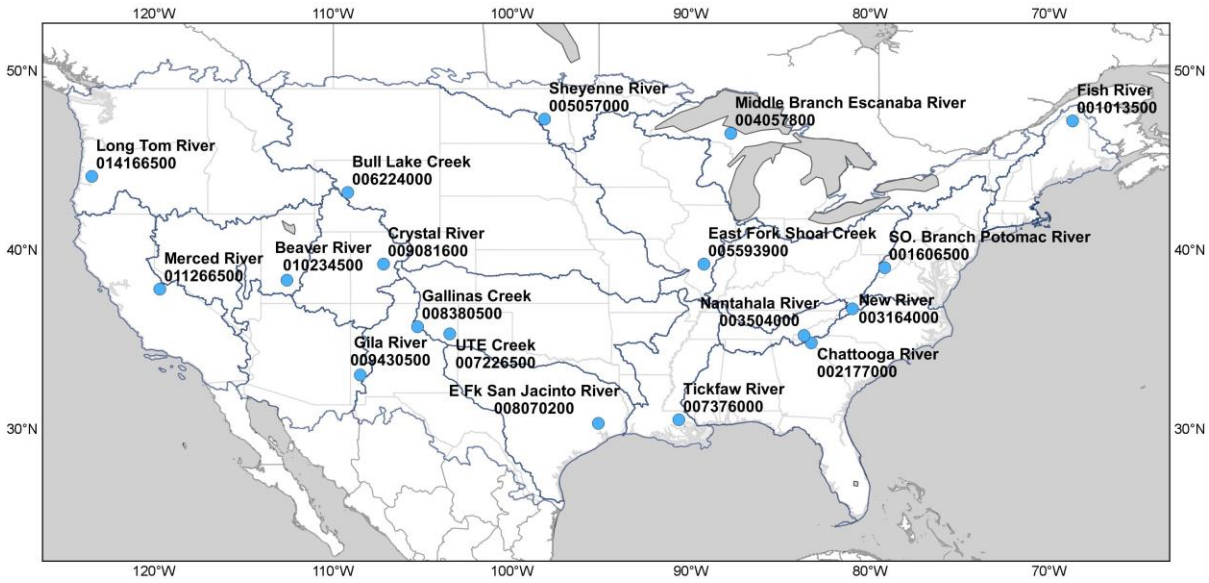
789

790 *Figure 2 Schematic of a skill surface plot. The y and the x axes display the SCF and the IHC skill,*

791 *respectively. They are expressed as a percentage of the climatological variance explained in*

792 *the respective predictability source. The blending weights, w_{SCF} and w_{IHC}, from which the skill*

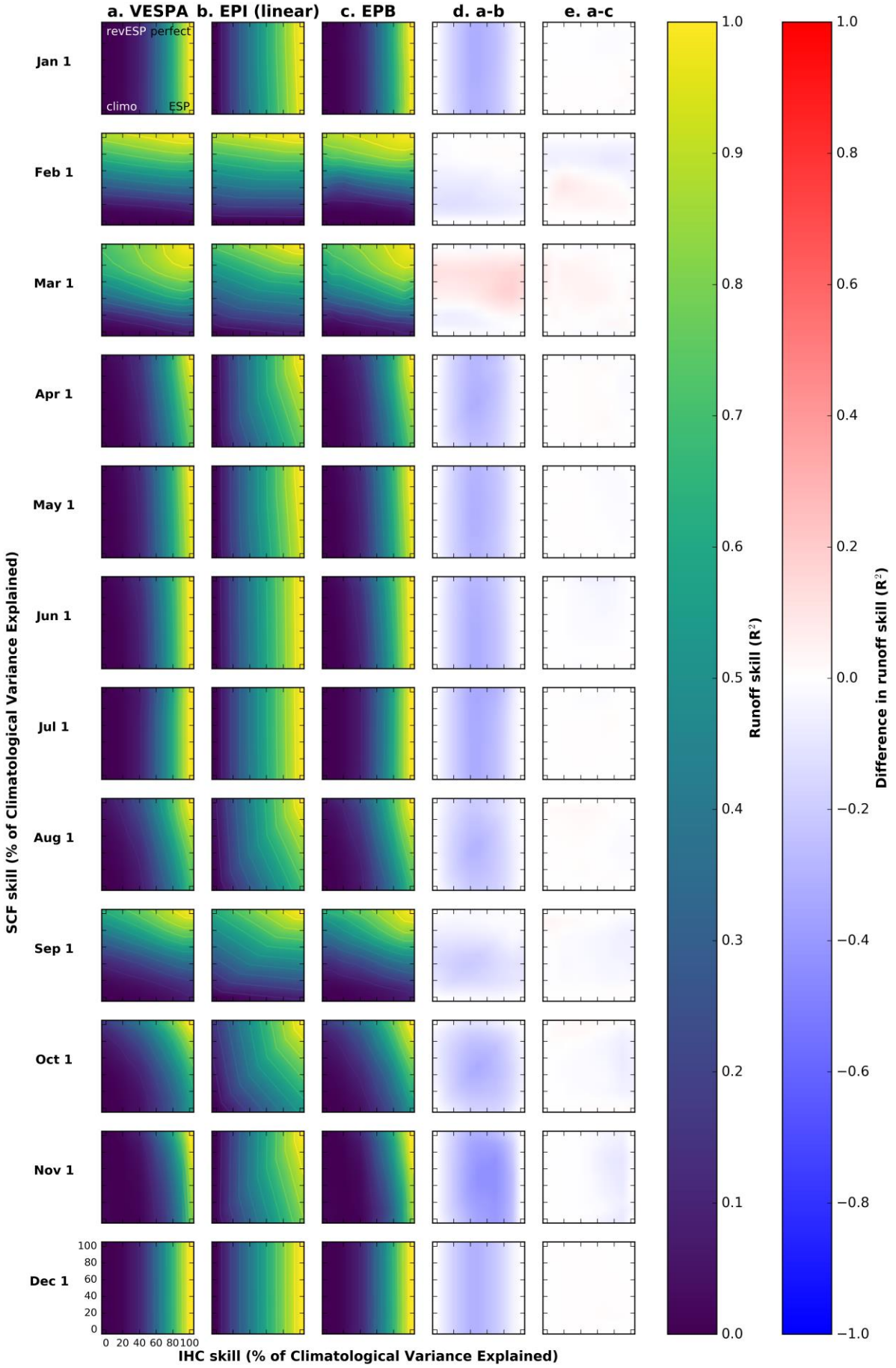
793 *values are derived are shown in square brackets in the figure.*



794

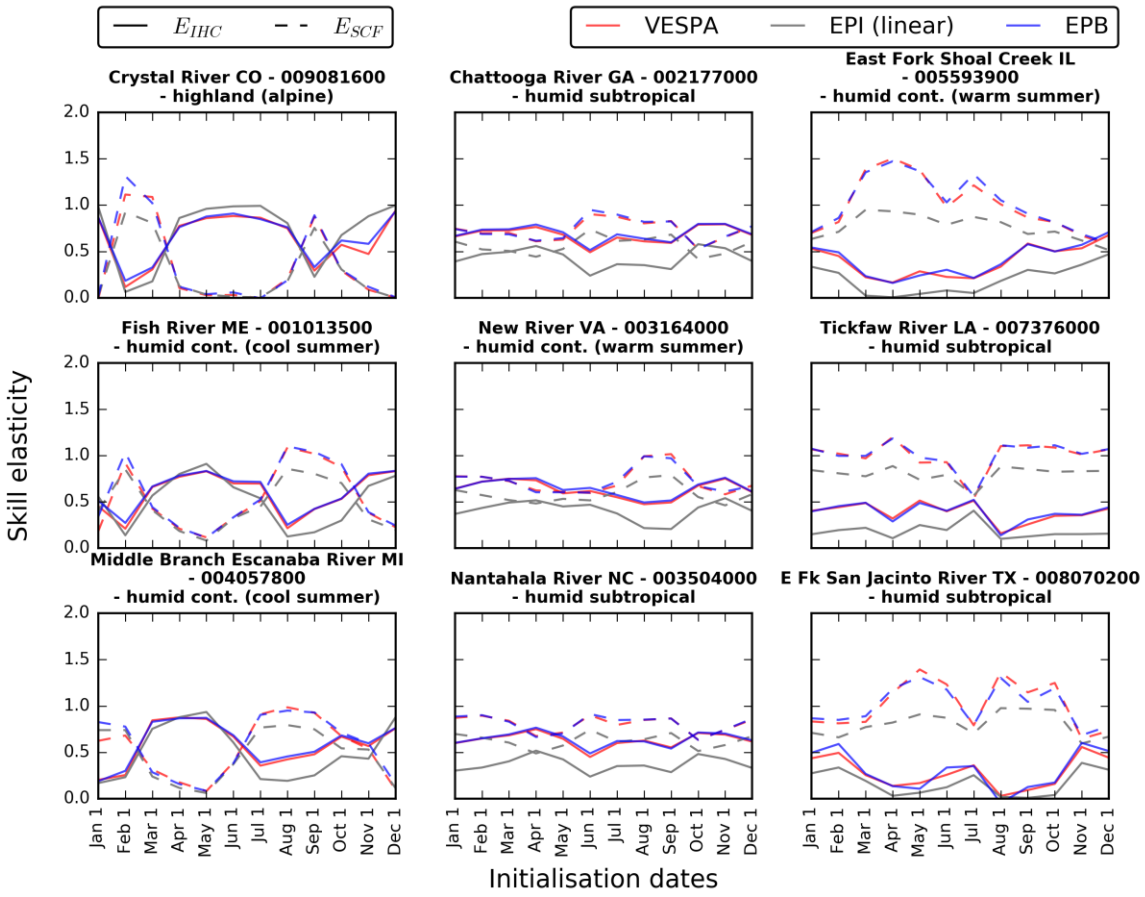
795 *Figure 3 Map of the 18 catchments of the CONUS selected for the analysis, and the HCDN*

796 *regions (dark blue outlines).*


a. VESPA
b. EPI (linear)
c. EPB
d. a-b
e. a-c
revESP
perfect
climo
ESP
Jan 1
Feb 1
Mar 1
Apr 1
May 1
Jun 1
Jul 1
Aug 1
Sep 1
Oct 1
Nov 1
Dec 1
SCF skill (% of Climatological Variance Explained)
IHC skill (% of Climatological Variance Explained)
0 20 40 60 80 100
Runoff skill (R^2)
Difference in runoff skill (R^2)
1.0
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
−0.2
−0.4
−0.6
−0.8
−1.0

797

Figure 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 1981-2010 for the Crystal River (CO; USGS gauge 009081600), with forecast initialisations on the first day of each month. Differences between the skill surface plots obtained for the d. VESPA and linear EPI methods and the e. VESPA and EPB methods are also shown.



813

814 *Figure 5 Streamflow forecast skill elasticities for the IHCs (i.e., E_{IHC}, solid line) and the SCFs (i.e.,*

815 *E_{SCF}, dashed line), calculated across a quadrant situated within the 3-month streamflow*

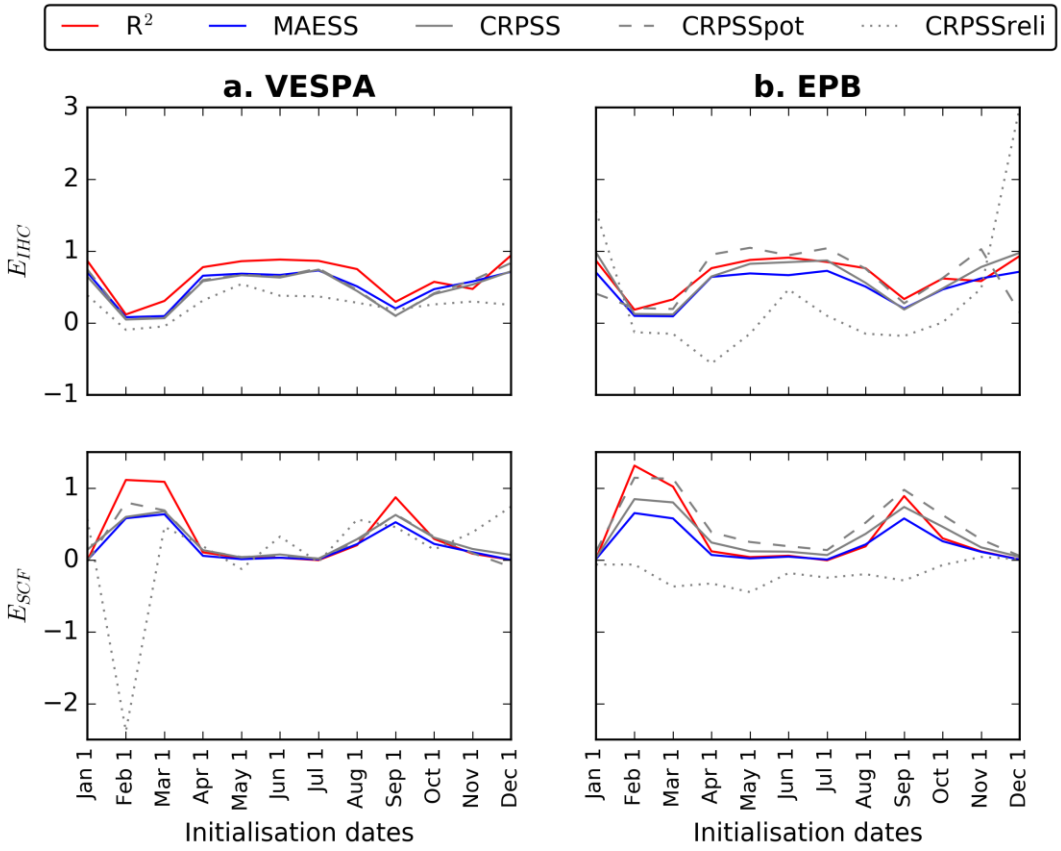
816 *forecast skill surface plots for the VESPA (in red), the linear EPI method (in grey) and the EPB*

817 *method (in blue; using Eq.* (4) and (5)*). Each plot shows the evolution of the IHC and SCF skill*

818 *elasticities with the initialisation date for a given catchment. The climatological regions of the*

819 *catchments are indicated in the plots' headings. The skill surface plots from which these skill*

820 *elasticities were calculated are presented in Figure 4 and Supplementary Figures 1 to 17.*



821

822 *Figure 6 Streamflow forecast skill elasticities for the IHCs (i.e., E_{IHC}, top two plots) and the SCFs*

823 *(i.e., E_{SCF}, bottom two plots) as a function of forecast initialisation dates, for hindcasts*

824 *produced from 1981-2010 for the Crystal River (CO; USGS gauge 009081600). These skill*

825 *elasticities were calculated across a quadrant situated within the 3-month streamflow*

826 *forecast skill surface plots (from Eq.* (4) and (5)*) for several verification scores (the R^2 in red,*

827 *the MAE skill score [MAESS] in blue, the CRPSS in grey solid line, the potential CRPSS*

828 *[CRPSSpot] in grey dashed line and the CRPSS reliability [CRPSSreli] in grey dotted line). The*

829 *results are shown for a. the VESPA approach (two leftmost plots) and b. the EPB method (two*

830 *rightmost plots).*