University Library

MINERVA ACCESS

A gateway to Melbourne's research publications

Minerva Access is the Institutional Repository of The University of Melbourne

Author/s:
Bennett, JC;Wang, QJ;Robertson, DE;Bridgart, R;Lerat, J;Li, M;Michael, K

Title:
An error model for long-range ensemble forecasts of ephemeral rivers

Date:
2021-05-01

Citation:
Bennett, J. C., Wang, Q. J., Robertson, D. E., Bridgart, R., Lerat, J., Li, M. & Michael, K. (2021). An error model for long-range ensemble forecasts of ephemeral rivers. Advances in Water Resources, 151, https://doi.org/10.1016/j.advwatres.2021.103891.

Persistent Link:
https://hdl.handle.net/11343/354213

An error model for long-range ensemble forecasts of ephemeral rivers

James C. Bennett[1,2*], Q.J. Wang[3], David E. Robertson[1], Robert Bridgart[1], Julien Lerat[5], Ming Li[4], Kelvin Michael[2]

[1]CSIRO Land & Water, Clayton, Victoria, Australia;

[2]Institute for Marine and Antarctic Studies, University of Tasmania, Battery Point, Tasmania, Australia.

[3]Department of Infrastructure Engineering, University of Melbourne, Parkville, Victoria, Australia;

[4]CSIRO Data61, Kensington, Western Australia, Australia.

[5]Bureau of Meteorology, Canberra, Australian Capital Territory, Australia.

[*]Correspondence to: james.bennett@csiro.au

Key points

- No existing long-range forecasting methods produce reliable ensembles in highly ephemeral streams.
- This is a new approach to handling zero values in both observations and simulations.
- Forecasts are reliable to 12 months for even highly ephemeral streams.

Abstract

Few ensemble streamflow forecasting systems are designed to operate for ephemeral rivers. In this study, we revise our error model for generating Forecast Guided Stochastic Scenarios (FoGSS) to produce statistically reliable long-range (12-month) forecasts for ephemeral rivers. FoGSS features an error model with four stages: data transformation, bias-correction, an autoregressive error model and the statistical distribution of residuals. We revise the fourth stage of FoGSS with a parameter estimation method that uses data censoring to account for zero values in both observations and forecasts. This allows FoGSS to produce statistically reliable ensemble forecasts in even highly ephemeral streams (with >50% zero flows). We apply FoGSS to conventional ensemble hydrological prediction (ESP) forecasts for 50 Australian catchments, including 26 ephemeral rivers. We show that FoGSS improves the accuracy of ESP forecasts at short lead times, while at long lead times FoGSS forecasts transition to climatology-like forecasts. FoGSS forecasts are reliable in ensemble spread at individual lead times and for volumes aggregated over lead times, even in highly ephemeral rivers. FoGSS forecasts pave the way for operational long-range forecasts in ephemeral rivers, meeting a key need for improved water management.

Keywords: ephemeral rivers; dryland streams; long-range forecasting; ESP; FoGSS; ensemble prediction.

1 Introduction

Ephemeral rivers pose particular problems for ensemble streamflow forecasting. They often exhibit highly non-linear responses in runoff to rainfall, making them difficult to model (e.g., Costigan et al., 2017). At the same time, statistical treatments that are necessary to generate reliable ensembles are complicated by the presence of zero values (McInerney et al., 2019; Smith et al., 2015; Smith et al., 2010). Perhaps as a consequence, few streamflow forecasting systems are designed for ephemeral rivers. This is despite the clear need for methods that can generate skillful and reliable streamflow forecasts for ephemeral rivers: ephemeral rivers drain close to half the earth's surface (Datry et al., 2017; Tooth, 2000), providing water vital to ecosystems and humans. In Australia, the need for streamflow forecasts in ephemeral rivers is acute. Ephemeral rivers are a crucial source of water for irrigated agriculture over much of Australia, for example in the northern and western Murray-Darling Basin, a crucial region in Australia's most productive agricultural basin. Balancing the competing needs of irrigators and ecosystems in water-stressed systems such as the Murray-Darling basin is one of Australia's most pressing water management challenges (e.g. Grafton and Wheeler, 2018). Further, prospective agricultural development in monsoonal northern Australia (e.g. Petheram et al., 2018) will rely heavily on ephemeral streams. Streamflow forecasts - in particular long-range (12-month) forecasts that assist in water allocation decisions (e.g. Kaune et al., 2020) - are likely to be beneficial in these and other Australian drylands.

Long-range forecasts of streamflow are often highly uncertain, and thus require ensemble forecasting methods. To be most useful to water managers, long-range ensemble forecasts should:

1. Be in the form of hydrographs at the monthly time step. This enables water managers to i) consider forecasts of individual months and ii) accumulate forecasts for longer periods (e.g. 6-month totals) to understand forecasts at a range of temporal scales.
2. Be statistically reliable, both at individual months and for accumulated volumes, at all lead times and at all times of year – including for ephemeral rivers.
3. Be as accurate as possible when skill is available, and never less accurate than simple climatology forecasts (a property known as 'coherence', after Krzysztofowicz, 1999).

We review existing approaches to ensemble prediction in light of these properties, with particular reference to ephemeral rivers.

Among the longest standing methods to produce long-range streamflow forecasts are simple regression models that describe a relationship between a predictor (e.g. an estimate of moisture stored in soils) and future streamflow. These methods have been used for decades in the United States (Pagano et al., 2009). This idea has been extended to more complex Bayesian models, including the Bayesian Joint Probability (BJP) modelling approach (Wang and Robertson, 2011; Wang et al., 2009) that underpins Australia's national seasonal streamflow forecasting service provided by the Bureau of Meteorology (BoM; http://www.bom.gov.au/water/ssf/). The BJP formally accounts for heteroscedasticity (the inability to assign a single variance to streamflow), parameter uncertainty and, crucially, zero values, enabling it to produce statistically reliable ensemble forecasts for ephemeral rivers. As with many statistical models, forecasts issued with the BJP are for a single lead time (total streamflow for the coming 1, 2 or 3 months), rather than hydrographs that connect streamflow forecasts at multiple lead times. While it is possible to break down 3-month streamflow totals into three 1-month forecasts with the BJP (Zhao et al., 2016), it is difficult to construct long-range hydrographs from simple predictor/predictand relationships. This is in large part because it is difficult to model the temporal properties of streamflow without some form of state variable(s) (e.g., states that mimic soil moisture stores). If temporal properties of individual ensemble members are not realistic, then aggregations of the ensemble will not be reliable - even if predictions at individual lead times are reliable (e.g. Shrestha et al., 2015).

Generating hydrographs is more easily performed with hydrological models. The practice of using hydrological models in seasonal streamflow prediction also has a long history: ensemble streamflow prediction (ESP) methods have been used since the 1970s (Day, 1985). ESP works by initialising a hydrological model with observed meteorological forcings (e.g. rainfall, potential evaporation) and then generating a forecast by running the model with an ensemble of these meteorological forcings taken from the historical record. An ensemble of historical forcings is simple to collate and inherently reliable (assuming a stationary climate). Forecast skill in ESP forecasts derives entirely from initial hydrological conditions (soil moisture, etc.) – i.e., forcings give no information about future meteorological conditions. Conversely, the ensemble spread in ESP forecasts derives entirely from the ensemble of forcings. ESP forecasts ignore the large uncertainties that arise in the conversion of rainfall to runoff, resulting in ensemble streamflow forecasts that are often over-confident and thus unreliable (Wood and Schaake, 2008).

More modern forecasting systems combine ensemble climate forecasts with hydrological models (e.g. Arnal et al., 2018; Crochemore et al., 2016). Without statistical processing, uncertainties in climate forecasts and the simulation of hydrological processes are often incorrectly specified – including in highly sophisticated ensemble climate prediction systems (Yuan et al., 2015; Zhao et al., 2017). That is, the ensembles they produce are unreliable; most usually they produce ensemble spread that is too narrow. The most common method to enforce reliable ensembles from such streamflow forecasting systems is to statistically calibrate them with methods analogous to the Model Output Statistics (MOS) approaches long used in meteorology (Hemri and Klein, 2017; Pokhrel et al., 2013; Verkade et al., 2017; Woldemeskel et al., 2018; Wood and Schaake, 2008). However, MOS approaches suffer from the same limitation described for the statistical forecasting methods above: calibration is applied separately at each lead time, completely disrupting temporal relationships of ensemble members and thus making it difficult to issue forecasts in the form of hydrographs.

An alternative approach is to separate forecast uncertainties into those related to climate forcings and those related to hydrological modelling. A range of methods already exists to effectively calibrate climate forecasts (e.g. Strazzo et al., 2019; Wang et al., 2019). This leaves the challenge of specifying hydrological uncertainties, which is the subject of this study. We address this challenge with error modelling. Error models have the crucial advantage over MOS methods of preserving the temporal sequences of forecast hydrographs. Most hydrological error models are developed to predict only one lead time in advance, but it is possible to propagate uncertainty through multiple lead times (Seo et al., 2006). We have previously developed the FoGSS (Forecast Guided Stochastic Scenarios) error model to generate reliable, long-range (to 12 months) ensemble forecasts in the form of hydrographs at the monthly time step (Bennett et al., 2016b; Bennett et al., 2017; Li et al., 2013). FoGSS is based on three principles:

i) Key error model parameters are adjusted by calendar month (Li et al., 2013; Liu et al., 2020)
ii) The error model is comprised of several independent stages, each addressing a particular aspect of model errors, rather than a single complex model (Bennett et al., 2016b; Bennett et al., 2017; Li et al., 2015; Li et al., 2016; Li et al., 2017).
iii) Error model parameters that are estimated for lead one predictions are used to calibrate forecasts to long lead times (Bennett et al., 2016b; Bennett et al., 2017; Li et al., 2017)

We have shown that FoGSS is able to generate reliable streamflow forecasts to lead times of 12 months across a wide variety of perennial catchments (Bennett et al., 2016b; Bennett et al., 2017). FoGSS reduces forecast error with an autoregressive model, particularly at shorter lead times. Further, the forecasts it produces are 'coherent': that is, they are always at least as skillful as climatology forecasts, even at long lead times. However, the success of FoGSS in ephemeral catchments was mixed, and depended on the degree of ephemerality. It worked as expected in moderately ephemeral catchments (defined as never having any calendar month with >50% zeros), but performed poorly in highly ephemeral catchments (where some months had >50% zeros) (Bennett et al., 2017). FoGSS was unable to generate reliable ensembles in highly ephemeral rivers, sometimes resulting in negative forecast skill with respect to climatology.

Accounting for zeros in hydrological error models is both technically challenging and essential. Smith et al. (2010) showed that for parameter estimation methods that use a likelihood, it is not possible to correctly optimise error model parameters for ephemeral rivers without explicitly accounting for zeros. They developed a likelihood that accounted for zero values by treating residuals as a mixed discrete-continuous distribution, conditioned on observations. They later complemented this work by treating residuals as autocorrelated (Ammann et al., 2019; Smith et al., 2015). Other error models, including FoGSS, use data censoring rather than a mixed discrete-continuous distribution to treat the presence of zeros in observations (Li et al., 2016; Li et al., 2017; McInerney et al., 2019). Error models that explicitly handle the presence of zeros in observations generally perform well for modelling errors in cases where i) catchments are not more than moderately ephemeral and ii) hydrological simulations do not equal zero. However, because these error models treat residuals as symmetrically distributed around the hydrological model simulation (usually after transformation to normalise data), they cannot generate >50% zeros (Wang et al., 2020). That is, error models that handle zeros only in observations are structurally incapable of reliably simulating errors in highly ephemeral catchments, as we describe in detail in Section 2.2. For this reason, FoGSS and similar models are unsuitable for generating reliable predictions in highly ephemeral catchments.

In a recent paper, we addressed the problems of error models in highly ephemeral streams (Wang et al., 2020). We used data censoring to establish a likelihood that applies censoring to hydrological simulations as well as observations. This likelihood was applied to a simple Gaussian error model that assumed uncorrelated residuals and produced reliable simulations

even in highly ephemeral rivers. However, the aims of Wang et al. (2020) were to improve hydrological simulations and to correctly specify uncertainty in highly ephemeral streams, not to produce forecasts at multiple lead times.

In the present study, we adapt the likelihood proposed by Wang et al. (2020) to work with the FoGSS error model to produce long-range streamflow forecasts in ephemeral rivers. We test the revised FoGSS error model on ESP forecasts generated with an experimental hydrological-model based streamflow forecasting system developed by the Bureau of Meteorology (Woldemeskel et al., 2018). To demonstrate the FoGSS error model's more general applicability, we test the forecasts on perennial, moderately ephemeral and highly ephemeral rivers.

The paper is structured as follows. We give a brief overview of the selected terms used in this paper in Section 1.1. The FoGSS error model is described in Section 2, dividing the existing components (Section 2.1) from the changes made for ephemeral rivers (Section 2.2). Section 3 describes the ESP forecasting system to which we apply FoGSS, and the process fo generating FoGSS forecasts is described in Section 4. Catchments and data are described in Section 5. Forecast verification methods, including our cross-validation scheme, are described in Section 6, and the results of our experiments are presented in Section 7. We discuss our findings in Section 8 and summarise and conclude the study in Section 9.

1.1. Terminology

Several terms used in this paper are used in specific ways or are not widely known, and we define these here to assist the reader.

FoGSS versions: We designate the previous FoGSS error model (Bennett et al., 2017) as 'o-censored', to indicate that it only applies censoring to observations, and the new FoGSS presented in this paper as 'os-censored', to indicate that it applies censoring to both observations and simulations. This follows the terminology used by Wang et al. (2020).

Lead times: FoGSS produces forecasts monthly forecasts to a horizon of 12 months into the future. We refer to each forecast lead time as 'lead 1', 'lead 2',…, 'lead 12'. A lead 1 forecast is a forecast for the next month: e.g., if a forecast is issued on 1 Feb 2000, the lead 1 forecast will predict streamflow for the period 2 Feb 2000-1 Mar 2000.

Hydrological simulations/forecasts: We designate hydrological 'simulations' as those generated by forcing the hydrological model with observed rainfall and potential evaporation.

205 'Forecasts' are generated by forcing an initialised hydrological model with forecast
206 rainfall/potential evaporation.

207 2 FoGSS error model

208 2.1 Previous FoGSS error model: o-censored

209 We briefly review the structure of o-censored FoGSS error model, which we have developed
210 previously through a series of papers (Bennett et al., 2016b; Bennett et al., 2017; Li et al.,
211 2013; Li et al., 2015). These papers may be consulted for further detail. Note that FoGSS is
212 designed to produce forecasts at the monthly time step. Forecasts are issued at the beginning
213 of each calendar month (i.e., 12 forecasts per year) to a lead time of 12 months.

214 2.1.1 Stage 1 – Data transformation

215 In previous studies using FoGSS (Bennett et al., 2016b; Bennett et al., 2017), Stage 1 has
216 encompassed the estimation of hydrological model parameters. However, it is possible to
217 apply the FoGSS error model to a hydrological model that is already calibrated. In this study
218 we wish to apply FoGSS to an existing forecasting setup (Section 3), and thus Stage 1
219 concerns only the transformation.

220 We use the log-sinh transformation (Wang et al., 2012) to allow us to treat our errors as
221 normally distributed. It is given by:

222 $$z(t) = Tf(q(t)) = \frac{1}{b}\log\left(\sinh\left(a + cq(t)b\right)\right) \quad (1)$$

223 where $q(t)$ is the observed or simulated monthly streamflow at time t, and a and b are
224 parameters. $c = \frac{5}{\max(\mathbf{q}_o)}$ is a standardisation constant where $\max(\mathbf{q}_o)$ is the maximum of the
225 time series of available streamflow observations. The standardisation constant allows a and b
226 to take comparable values across all catchments, simplifying the application of Bayesian
227 priors for parameter estimation (Appendix A). Note that when Eq 1 is reversed (to back-
228 transform z to q), any values of $z(t) < Tf(0)$ (the transformed value of zero) are first forced
229 to $z(t) = Tf(0)$ before back-transformation, to ensure streamflow values cannot be negative.

230 Transformation parameters are estimated only from observations using a Maximum A
231 Posteriori (MAP) estimation (Section 2.2.2), and these parameters are applied to both
232 observed and simulated streamflow.

2.1.2 Stage 2 – Bias-correction

A bias-correction is applied to transformed simulations at each calendar month:

$$z_2(t) = d(i)\,z_1(t) + \mu(i) \quad (2)$$

where $i = 1,2,\ldots,12$ is the calendar month corresponding to t, $z_1(t)$ is the log-sinh transformed simulation (Stage 1), and $d(i)$ and $\mu(i)$ are parameters that vary by month. We limit d to $0 \le d < 2$. Values less than zero imply negative correlations (i.e., very poor hydrological model performance), meaning that it is more sensible to ignore simulations ($d = 0$). The upper limit is arbitrary, and avoids overly large corrections. The bias-correction parameters are estimated using a method we term 'Least Squares after Transformation' (LST), detailed in Section 2.2.2.

The bias-correction performs two important tasks. First, it allows us to treat our residuals as normally distributed with a mean of zero (see Stage 4). Second, as $d(i)$ approaches zero (which can occur when hydrological simulations are very poor), the bias-correction tends to $z_2(t) \approx \mu(i)$; that is, it returns a constant, akin to climatology. This is similar in concept to the statistical calibration of forecasts with MOS methods. As noted in the introduction, MOS methods apply separate regressions to each lead time, breaking the temporal sequence of ensemble members. For our application (12-month forecasts issued 12 time per year) using MOS would mean applying a regression at each forecast issue time and each lead time ($12 \times 12 = 144$ regressions). By contrast, in FoGSS we estimate the bias-correction for each calendar month at lead 1, and apply the bias-correction to that calendar month at all lead times (at total of 12 regressions). This allows us to preserve the temporal sequence from the hydrological model in the forecast (Section **Error! Reference source not found.**). We have shown in previous papers that this method acts like a meteorological calibration at all lead times, returning climatology forecasts when predictions are not skilful (Bennett et al., 2016b; Bennett et al., 2017).

2.1.3 Stage 3 – Autoregressive updating

In Stage 3 we apply a first-order autoregressive (AR1) that follows the general form

$$z_3(t) = z_2(t) + \rho(i)\left(z_o(t-1) - z_2(t-1)\right) \quad (3)$$

where z_2 is the transformed bias-corrected simulation from Eq 2, $z_o(t-1) = Tf\left(q_o(t-1)\right)$ is the log-sinh transformed value of observed flow, and $\rho(i)$ is a parameter that varies by calendar month, which can take values between 0 and 1. The transformation (Eq 1) can amplify values of $z_3(t)$ to be unrealistically large after back transformation and to avoid this problem, we restrict $z_3(t)$ (Li et al., 2015):

$$z_3(t) = \begin{cases} Tf\left(\min\left(q_3(t), q_2(t) + \left(q_o(t-1) - q_2(t-1)\right)\right)\right) & \text{when } q_o(t-1) \geq q_2(t-1) \\ Tf\left(\max\left(q_3(t), q_2(t) + \left(q_o(t-1) - q_2(t-1)\right)\right)\right) & \text{when } q_o(t-1) < q_2(t-1) \end{cases} \quad (4)$$

where $q_2(t) = Tf^{-1}\left(z_2(t)\right)$ and $q_3(t) = Tf^{-1}\left(z_3(t)\right)$ are the back-transformed values of $z_2(t)$ and $z_3(t)$. In other words, we limit the size of the update to the error in the original domain. The AR1 model reduces errors and plays a pivotal role in propagating uncertainty through multiple lead times, as described in Section **Error! Reference source not found.**. As with the bias-correction parameters, the AR1 parameters are estimated using a method we term 'Least Squares after Transformation' (LST), detailed in Section 2.2.2.

2.2 Revising FoGSS for ephemeral rivers: os-censored

2.2.1 Stage 4 - residual modelling

Previous (o-censored) versions of FoGSS described only 3 stages, with the statistical modelling of residuals included in Stage 3. For os-censoring, we separate out the statistical model of residuals into Stage 4 to highlight the improvements to FoGSS for ephemeral catchments.

As with previous incarnations of FoGSS, we assume residuals follow a normal distribution:

$$\begin{aligned} z_4(t) &= z_3(t) + \varepsilon(t) \\ \varepsilon(t) &\sim N\left(0, \sigma^2(i)\right) \end{aligned} \quad (5)$$

where z_3 is the transformed, bias-corrected and updated simulation from Eq 4, and $\sigma^2(i)$ is the standard deviation of residuals, varied by calendar month. To generate predictive uncertainty, we revise FoGSS to follow Wang et al. (2020)'s use of data censoring of both observations and simulations (os-censoring). Denote a censoring threshold of $q_c = 0$, and let $z_c = Tf(q_c)$ be the transformed censoring threshold. If $z_3(t) \leq z_c$, a new value is assigned to $z_3(t)$ by

$$z_3(t) = \Phi^{-1}\left(\Phi\left(z_C \mid m_3(i), s_3(i)^2\right) \times r_O(t)\right)$$
$$r_0(t) \sim U(0,1) \quad (6)$$

where Φ is the normal cumulative probability, $m_3(i)$ and $s_3(i)$ are the mean and standard deviation of $z_3(t)$ for the calendar month i corresponding to t, and $U(0,1)$ is a uniform distribution (see Section 2.2.2 for the estimation of these parameters). We then add noise to $z_3(t)$ to give $z_4(t)$ by sampling from Eq (5). Note that when we generate a forecast, Eq (5) and Eq (6) are applied at each lead time through a process we call *stochastic updating*. We describe the process of generating a forecast, including stochastic updating, in Section **Error! Reference source not found.**.

Eq (6) is a crucial advance over the o-censored FoGSS as it allows the error model to produce >50% zero values, which we show with a schematic in Fig 1. In conventional error models, $z_3(t)$ can only take values $z_3(t) \geq z_c$. If we assume a symmetrical error distribution (e.g., Eq 5) at most 50% of $z_4(t)$ can be less than or equal to z_c. (Remember that any values of $z_4(t) < z_c$ are forced to $z_4(t) = z_c$ before back-transformation.) In the os-censored FoGSS error model, Eq (6) re-assigns values of $z_3(t) = z_c$ to $z_3(t) \leq z_c$. When noise is added according to Eq (5), the predictive distribution can now produce >50% zeros after back-transformation. This enables reliable predictive uncertainty to be generated for highly ephemeral streams.

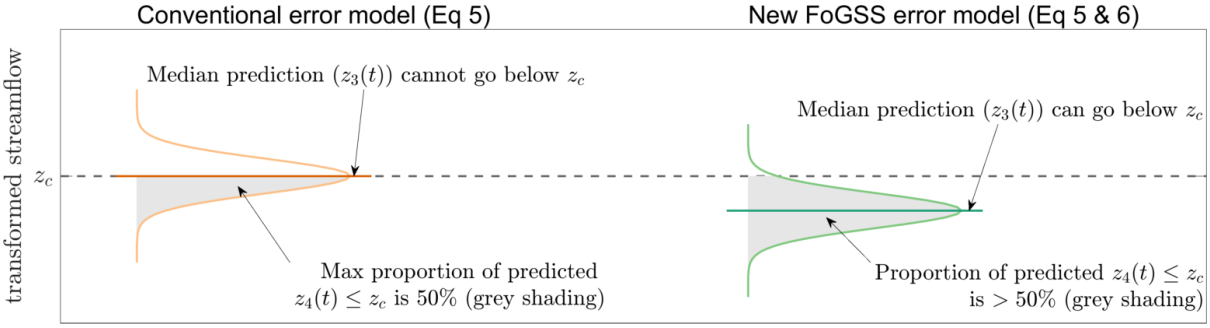


Figure 1 Schematic of FoGSS error model compared to conventional error models, adapted from Wang et al. (2020). Left hand example shows a conventional error model (Eq 5), where the median transformed simulation $z_3(t) \geq z_c$, and thus at most 50% of $z_4(t)$ values can be less than or equal to z_c. Right hand shows the FoGSS error model, where Eq. 6 re-assigns a value of $z_3(t) \leq z_c$, allowing >50% of $z_4(t)$ to be less than or equal to z_c.

Wang et al. (2020) noted that because of the way many hydrological models are structured, their simulations can never reach exactly zero. The GR4J model we use in this study can

suffer from this issue. In such cases $z_3(t)$ is always greater than $Tf(0)$, meaning Eq (6) is never enacted if $q_c = 0$. This makes reliable predictive uncertainty impossible to generate for streams with >50% zeros. Accordingly, Wang et al. (2020) suggested using a censoring threshold slightly above zero to ensure Eq (6) is enacted. In FoGSS, however, this is unnecessary. The bias-correction (Stage 2) and AR1 model (Stage 3) can correct simulations to zero, ensuring Eq (6) is enacted frequently with a censoring threshold of $q_c = 0$.

Stage 4 parameters are estimated by Maximum Likelihood

2.2.2 Revised parameter inference for os-censored FoGSS

FoGSS error model parameters are estimated at lead one, as with a conventional hydrological model calibration. That is, one set of parameters is applied to all lead times when forecasts are generated. Error model parameters are inferred independently at each stage. In all cases, a numerical search method is used to find optimal parameters (Duan et al., 1993). We provide a basic overview of the parameter estimation procedure here. A more technical description, including equations and priors, is given in Appendix A.

Stage 1

As noted in Section **Error! Reference source not found.**, in previous versions of FoGSS we have included the estimation of hydrological model parameters with the error model at Stage 1. In this study, we apply FoGSS to an existing forecasting setup, where the hydrological model is already calibrated (Section 3.1). Thus in Stage 1 only the transformation parameters are estimated.

The Stage 1 transformation parameters are estimated by *Maximum A Posteriori* (MAP) estimation, detailed in Appendix A. Note that Stage 1 applies a single set of transformation parameters for all data, rather than varying transformations by month. This means that all forecasts and observations take a common range of values in the transform domain, greatly simplifying the AR1 modelling in Stage 3.

Stage 2 and Stage 3

It is possible to use maximum likelihood estimation (MLE) for stages 2 and 3, using a likelihood similar to that described for Stage 4 (below, and Appendix A). However, the likelihood is quite computationally intensive. Further, we do not require an estimate of residual uncertainty at stages 2 or 3, as that is the role of Stage 4. We can therefore use a

simple short-cut to reduce computation at Stage 2 and Stage 3. Rather than the likelihood, we use least squares of residuals on transformed data (abbreviated here to LST) to achieve similar results (Chatterjee and McLeish, 1986; Li et al., 2020). As we normalize our data at Stage 1, the assumption of normal residuals that underlies least squares regression will be satisfied. LST reduces computation considerably. For an observed time series of length $t = 1, 2, \ldots T$, the LST objective minimizes

$$L = \sum_{t=1}^{T} \left(\max\left(z_o(t), z_c\right) - \max\left(z_S(t), z_c\right) \right)^2 \quad (7)$$

where $z_S(t)$ is the transformed simulation at stage $S = 2$ or $S = 3$.

Stage 4

Stage 4 parameters are estimated by MLE, using the likelihood proposed by Wang et al. (2020). The os-censored likelihood handles four cases, depending on whether the simulation and/or observation are equal to zero:

- case = 1: observed and simulated flow are both greater than zero, i.e. $q_o(t) > 0$ and $q_3(t) > 0$

$$p\left(q_4(t) = q_o(t) \mid q_3(t)\right) \propto N\left(z_o(t) \mid z_3(t), \sigma^2(i)\right) \quad (8)$$

- case = 2: observed flow is zero and simulated flow is greater than zero, i.e. $q_o(t) = 0$ and $q_3(t) > 0$

$$p\left(q_4(t) = q_o(t) = 0 \middle| q_3(t)\right) \propto \Phi\left(z_c \middle| z_3(t), \sigma^2(i)\right) \quad (9)$$

- case = 3: observed flow is greater than zero and simulated flow equals zero, i.e. $q_o(t) > 0$ and $q_3(t) = 0$

$$p\left(q_4(t) = q_o(t) \middle| q_3(t) = 0\right) \propto \frac{N\left(z_c \middle| m_3(i), s_3^2(i) + \sigma^2(i)\right) \Phi\left(z_c \middle| \frac{s_3^2(i) z_3(t) + \sigma^2(i) m_3(i)}{s_3^2(i) + \sigma^2(i)}, \frac{\sigma^2(i) s_3^2(i)}{s_3^2(i) + \sigma^2(i)}\right)}{\Phi\left(z_c \middle| m_3(i), s_3^2(i)\right)} \quad (10)$$

- case = 4: observed and simulated flow are both zero, i.e. $q_o(t) = 0$ and $q_3(t) = 0$

$$p\left(q_4(t) = q_o(t) = 0 \middle| q_3(t) = 0\right) \propto \frac{\int_{-\infty}^{z_c} \Phi\left(z_c \middle| z_3(t), \sigma^2\right) N\left(z_3(t) \middle| m_3(i), s_3^2(i)\right) dz_3(t)}{\Phi\left(z_3(t) \middle| m_3(i), s_3^2(i)\right)} \quad (11)$$

where $q_3(t)$ and $z_3(t) = Tf\left(q_3(t)\right)$ are hydrological model simulations at Stage 3 before and after transformation, and $q_o(t)$ and $z_o(t) = Tf\left(q_o(t)\right)$ are observations before and after transformation. $z_c = Tf(q_c)$ is the transformed censoring threshold, $m_3(i)$ and $s_3(i)$ are the mean and standard

deviation of $\mathbf{z}_3$ and Φ is the normal cumulative probability. Note that each of $m_3(i)$, $s_3^2(i)$, and $\sigma^2(i)$ take different values for different calendar months (see Eq 5 and Eq 6). $m_3(i)$ and $s_3^2(i)$ are estimated from $q_3(t)$ simulations generated by applying the first 3 stages of FoGSS to the hydrological model simulation (see Appendix A).

We denote the residual distribution parameters estimated at Stage 4 by $\theta_4 = \left\{\sigma^2(i):\ i=1,2,\ldots,12\right\}$. The likelihood is given by:

$$L(\theta_4) \propto \prod_{t:\ case=1} p\left(q_4(t)=q_o(t)\,|\,q_3(t)\right) \prod_{t:\ case=2} p\left(q_4(t)=q_o(t)=0\,|q_3(t)\right) \prod_{t:\ case=3} p\left(q_4(t)=q_o(t)\,|q_3(t)=0\right) \prod_{t:\ case=4} p\left(q_4(t)=q_o(t)=0\,|q_3(t)=0\right) \quad (12)$$

where the cases are described by equations (8)-(11). Eq (10) and Eq (11) make this likelihood unique in the literature: they treat simulated values as censored data. As described in Section 2.2.1, the treatment of simulations as censored data at $q_3(t)=0$ enables Eq (6) to be enacted, making it possible to produce reliable predictive uncertainty in highly ephemeral streams (streams that have >50% zero flow in some months).

We have now described the structure of FoGSS and parameter estimation procedure for os-censoring. Before we describe the forecast generation procedure for FoGSS (Section 4) we describe the ESP forecasts that are the key input to the FoGSS error model (Section 3).

3 ESP forecasts

ESP forecasts are generated with the BoM's experimental 'dynamical' seasonal streamflow forecasting system (Feikema et al., 2018; Woldemeskel et al., 2018). This system is run at a daily time step, and in its usual configuration outputs are aggregated to 1-month and 3-month totals before MOS post-processing is applied. For our study, we aggregate daily forecasts to a monthly time step to produce streamflow forecasts as 12-month time series.

The dynamical forecast system usually uses calibrated climate forecasts as forcing, to lead times of ~90 days. While it is possible to generate calibrated daily forcings from climate prediction systems (Schepen et al., 2017), these methods have not been tested to the long lead times (365 days) we require in our study. ESP forcings are inherently reliable and simple to generate, and as this study focusses on the development of a hydrological error model, ESP forcings suffice. We discuss the prospects for using calibrated climate forecasts in combination with the FoGSS error model in Section 8.

3.1 Hydrological model

Hydrological modelling in the BoM's dynamical forecasting system is carried out with the daily GR4J model (Perrin et al., 2003), a simple four-parameter conceptual hydrological model that has performed strongly in Australian catchments in model intercomparison studies (Bennett et al., 2016a; Coron et al., 2012). It is forced by rainfall and potential evaporation (PE).

For each gauge, GR4J is calibrated by minimising the sum of squared errors computed on Box-Cox transformed flows (McInerney et al., 2017), following the BoM's existing practice. Calibration is carried out under cross-validation, as described in Section 6.1. To enable FoGSS parameters to be estimated, we run GR4J in simulation mode – i.e., forced with observed rainfall and PE - and aggregate daily GR4J simulations to monthly data to generate q_1. When running these simulations, we warm up GR4J for a minimum of 5 years (Section 3.3).

3.2 Rainfall and potential evaporation sampling

Rainfall and PE forcings in ESP forecasts are taken from historical observations. We sample ESP forcings with a leave-4-years-out cross-validation procedure (i.e., including the target year), which we illustrate by an example. To generate a forecast for January-Dec in 1980, we sample daily January-Dec rainfall and PE sequences from 1985, 1986, …, 2008. Each sequence is 1 year long. We construct sequences from 29 years of data (1980-2008). We choose 1980-2008 because this is the standard period used by the BoM to assess seasonal streamflow forecasting products. The choice of leaving out 4 years is less stringent than the cross-validation used for the hydrological component of the system (Section 5.1). Rainfall has far less interannual memory than streamflow, meaning that a less stringent cross-validation is acceptable. (The specific choice of leaving out four years was taken to generate a 25-member ensemble for both rainfall and PE. We ultimately generate 1000-member ensembles with FoGSS, and this process was simplified for an ESP ensemble of 25 members, because 1000 is evenly divisible by 25.)

PE and rainfall ensemble members are paired: i.e., if a rainfall ensemble member was sampled from 1985, the matching PE member is also from 1985. Rainfall and PE are often anticorrelated, so pairing rainfall and PE ensemble members is necessary to ensure realistic outputs from the hydrological model.

3.3 Generating ESP forecasts

We issue a forecast at the beginning of every calendar month. To generate a forecast, we initialise GR4J states by running it with observed forcings from 1/1/1975 onwards (a minimum of 5 years warmup for our reforecast period of 1980-2008). At the forecast issue time, the hydrological model is then forced with a single member of the rainfall/PE ensemble to produce a 1-year forecast of streamflow at the daily time step. This process is repeated for all 25 ensemble members to produce $\mathbf{q}_{F1}$. We then aggregate the daily forecasts to the monthly time step before applying the FoGSS error model.

4 Generating an os-censored FoGSS forecast

For a given forecast issue time, t, we assume we have available an initialised hydrological model simulation $q_1(t)$ at the forecast issue date, together with an observation $q_o(t)$. We also have available an uncorrected hydrological forecast $\mathbf{q}_{F1} = \{q_{F1}(t+1),\ldots,q_{F1}(t+\tau),\ldots,q_{F1}(t+12)\}$ for lead times $\tau = 1,2,\ldots,12$ (Section 3). We apply FoGSS to each ensemble member separately, where $\mathbf{q}_{F1}$ takes hydrological states from the hydrological simulation $q_1(t)$ and is then forced with a single member from the ensemble climate forecast. We apply the os-censored FoGSS error model to $\mathbf{q}_{F1}$ as follows:

1. $\mathbf{q}_{F1}$ is transformed (Stage 1 - Eq 1) and bias-corrected (Stage 2 - Eq 2) to produce $\mathbf{z}_{F2} = \{z_{F2}(t+1), z_{F2}(t+2),\ldots, z_{F2}(t+12)\}$. In the same way, $q_1(t)$ is transformed and bias-corrected to produce $z_2(t)$. Then $q_o(t)$ is transformed to produce $z_o(t)$.
2. For $\tau = 1$, we apply Eq (3) (Stage 3) as

$$z_{F3}(t+\tau) = z_{F2}(t+\tau) + \rho(i)\left(z_o(t) - z_2(t)\right) \tag{13}$$

   where i is the calendar month at time $t+\tau$. We also apply the restriction (Eq 4) but we omit this equation for brevity.

3. If $z_{F3}(t+\tau) \le z_c$ we assign a new value to $z_{F3}(t+\tau)$ by applying Eq (6):

$$\begin{aligned} z_{F3}(t+\tau) &= \Phi^{-1}\left(\Phi\left(0 \mid m_3(i), s_3(i)^2\right) \times r_O(t+\tau)\right) \\ r_0(t+\tau) &\sim U(0,1) \end{aligned} \tag{14}$$

4. We then draw a single realisation of noise according to Eq (5) (Stage 4):

454 $$z_{F4}(t+\tau) = z_{F3}(t+\tau) + \varepsilon(t+\tau)$$
$$\varepsilon(t+\tau) \sim N\left(0, \sigma^2(i)\right) \quad (15)$$

455 5. For lead times $\tau = 2,3,\ldots,12$ we no longer have an observation at the previous time
456 step. We resolve this by substituting z_{F4} from Eq (15) for z_o in Eq (13) to give:

457 $$z_{F3}(t+\tau) = z_{F2}(t+\tau) + \rho(i)\left(z_{F4}(t+\tau-1) - z_{F2}(t+\tau-1)\right) \quad (16)$$

458 where i is the calendar month at time $t+\tau$. (We again apply the restriction given by
459 Eq 4, again omitted for brevity.) We term the substitution of z_{F4} for z_o *stochastic*
460 *updating*, which we discuss in more detail below.

461 6. Apply Eq (14) and Eq (15) to $z_{F3}(t+\tau)$ from Step 5 (i.e., steps 3 and 4 are repeated for
462 values of $\tau > 1$).
463 7. Repeat steps 5 and 6 for each τ to produce the streamflow forecast ensemble member
464 $\mathbf{z}_{F4} = \left\{z_{F4}(t+1), ..., z_{F4}(t+\tau), \ldots, z_{F4}(t+12)\right\}$.
465 8. $\mathbf{z}_{F4}$ is converted to the original domain by reversing Eq (1):

466 $$\mathbf{q}_{F4} = Tf^{-1}\left(\max\left(\mathbf{z}_{F4}, z_c\right)\right) \quad (17)$$

467 We repeat steps 1-8 for each uncorrected ensemble member $\mathbf{q}_{F1}$ that has been generated with
468 in the ESP ensemble. Note that it is straightforward to generate large ensembles by repeating
469 steps 1-8 as many times as necessary for each $\mathbf{q}_{F1}$ - i.e., we can have many streamflow
470 forecast ensemble members ($\mathbf{q}_{F4}$) for each climate forecast ensemble member ($\mathbf{q}_{F1}$). As noted
471 in Section 3, the ESP forecasts have 25 ensemble members. Steps 1-8 are repeated 40 times
472 for each ensemble member to produce a 1000-member streamflow forecast ensemble.

473 Stochastic updating (Step 5) efficiently propagates uncertainty through the forecast (Bennett
474 et al., 2016b; Li et al., 2017). Estimating forecast uncertainty correctly at lead 1 is
475 straightforward, because FoGSS parameters are estimated at lead 1. However, uncertainty
476 should grow appropriately through lead times. The interaction of the random noise (Eq 15)
477 and the AR model (Eq 16) in stochastic updating allows uncertainty to grow through the
478 forecast. Uncertainty tends to grow rapidly at short lead times, but this growth slows and
479 ultimately stops as the influence of the AR model recedes at longer lead times. This is what
480 we are seeking in our uncertainty range: narrow uncertainty at short lead times when
481 forecasts are skillful, widening uncertainty as skill recedes with lead time, and an

approximately static uncertainty range (approaching a climatological distribution) when skill is absent. Stochastic updating is able to achieve this whilst retaining the temporal features of the hydrograph generated by GR4J.

5 Data

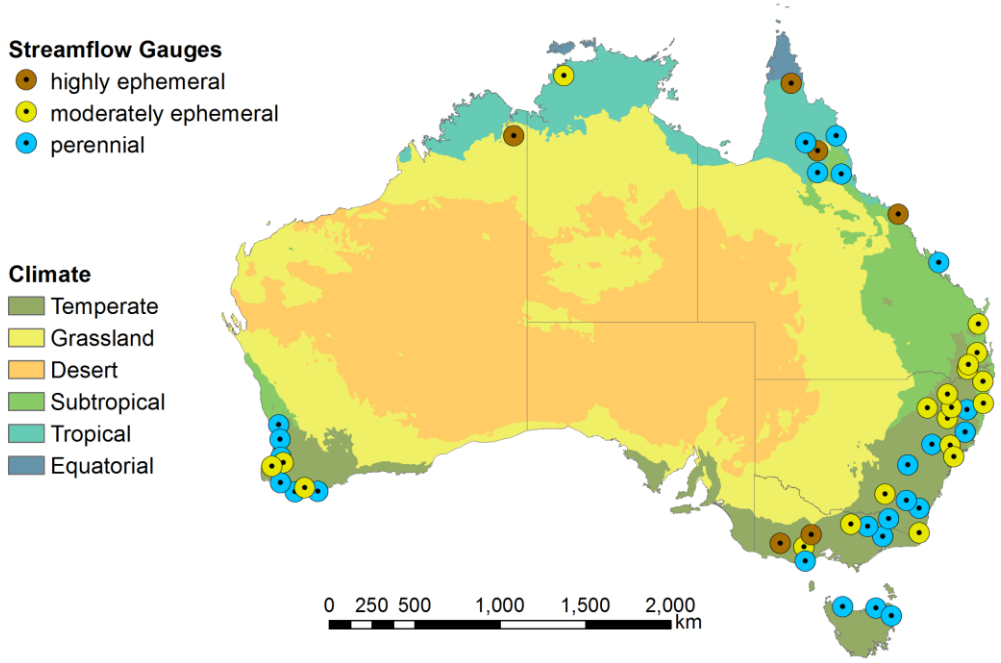
We assess ESP and FoGSS forecasts at 50 sites from a range of climates around Australia (Fig 2). We classify these sites into three categories of ephemerality based on observed flows at the monthly time step:

1) perennial (no month has >5% zero flows) – 24 sites
2) moderately ephemeral (some months have >5% zero flow, with no months >50% zero flow) – 20 sites
3) highly ephemeral (some months have >50% zero flow) – 6 sites

From these 50 sites, we select the 3 sites (one from each category) shown in Fig 3 to illustrate model performance in more detail. Note that the degree of ephemerality – in the sense of zero flows – is not replicated in the GR4J simulations. While it is technically possible for GR4J to produce zero flows for particular combinations of states/parameters, in practice this rarely occurs – i.e., like a lot of hydrological models, it tends not to produce zeros. For the 26 ephemeral catchments tested here, simulations were never zero.

Daily flow data are extracted from Water Data Online (http://www.bom.gov.au/waterdata/). Rainfall and potential evaporation (PE) are taken from the gridded AWAP data set (Australian Water Availability Project; Jones et al., 2009; http://www.csiro.au/awap/). AWAP produces daily estimates of rainfall interpolated from gauges to a ~5-km grid. PE estimates from AWAP are at a monthly time step; we disaggregate these to daily estimates by simple linear interpolation. Catchment estimates of rainfall and PE are calculated by areal averaging of AWAP grid cells that intersect with each catchment.

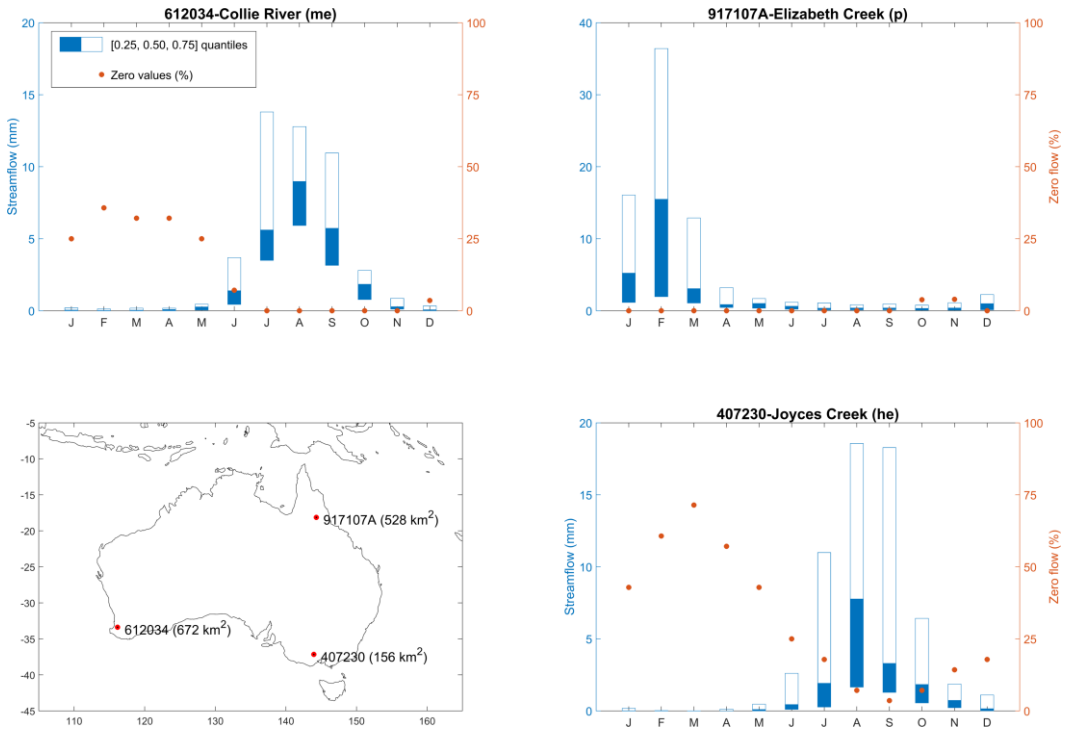
We use data for the period 01/01/1980-31/12/2012, which covers the 1980-2008 period used by the Bureau of Meteorology to assess all its seasonal streamflow forecasting products. We use the 01/01/1980-31/12/2012 period to estimate parameters (Section 2.2.2) and generate climatology reference forecasts (Section 6.2) under buffered cross-validation (Section 6.1) to ensure a consistent 4-year buffer for all years.



511

512 Figure 2 Location map of 50 sites used in this study, classified by degree of ephemerality (see
513 text for details). Climate zones are from Köppen climate classifications (Stern et al., 2000)

514



515

516 Figure 3 Exemplar catchments for moderately ephemeral (me), perennial (p) and highly
517 ephemeral (he) flow regimes

6 Forecast verification

6.1 Cross-validation scheme

All forecast verification is carried out under a buffered leave-one-year-out cross-validation scheme, with a buffer of 4 years. The cross-validation scheme is best described with an example. To attain GR4J and FoGSS parameters for 2000, we omit data from 2000 and the succeeding 4 years (2000-2004). The buffer is necessary to avoid informing the parameter estimation with rainfall information from the target year (2000), which can influence observed streamflow in subsequent years through catchment memory.

6.2 Verification scores

We focus on two probabilistic verification metrics: the probability integral transform (PIT) to measure ensemble reliability and the Continuous Ranked Probability Score (CRPS) to measure forecast accuracy (Gneiting and Katzfuss, 2014). Reliability measures the appropriateness of the ensemble spread: it should not be too wide (underconfident) nor too narrow (overconfident). CRPS measures both accuracy and reliability, but in the context of this study it is not strongly sensitive to reliability. This is because the benefits of the new (os-censored) FoGSS method are mainly that it produces more reliable forecasts at very low flow. Because CRPS is averaged over a range of forecasts, it tends to be more sensitive to errors in larger events than to the reliability of low flow months.

A PIT value is calculated for each forecast by

$$p(t)=\begin{cases} F\left(t,q_o(t)\right) & q_o(t)>0 \\ U(0,1)\times F(0) & q_o(t)=0 \end{cases} \quad (18)$$

where $F(t,[\ ])$ is the cumulative distribution function (CDF) of the forecast ensemble at time t. A set of forecasts at $t=1,2,\ldots,T$ is reliable if PIT values $\mathbf{p}=\{p(1),p(2),\ldots,p(T)\}$ are uniformly distributed. We check this by plotting PIT values against a standard uniform variate (we refer to these plots as 'PIT plots' for brevity). When PIT values follow the diagonal in PIT plots, the forecasting system is perfectly reliable. The treatment of PIT values at $q_o=0$ in Eq (18) is necessary to allow PIT values to follow a uniform distribution in the presence of zero values (Wang and Robertson, 2011). We refer to PIT values calculated when $q_o=0$ as 'pseudo-PIT' values.

Uniformity of PIT values can be summarised with the α-index (Renard et al., 2010). The α-index describes the tendency of PIT values to deviate from the diagonal in PIT plots:

$$\alpha = 1 - \frac{2}{N}\sum_{t=1}^{T}\left| p(t) - p_U(t) \right| \quad (19)$$

$p_U(t)$ is the theoretical value corresponding to $p(t)$. α ranges between 0 (unreliable) and 1 (perfectly reliable). Because it reduces PIT diagrams to a single value, α allows easy comparison between catchments and lead times.

CRPS is given by

$$C_M = \frac{1}{T}\sum_{t=1}^{T}\int_{-\infty}^{\infty}\left(F(t,x) - H\left(q_o(t) \le x\right)\right)^2 dx \quad (20)$$

where $F(t,[\ ])$ is the CDF of the forecast ensemble at time t, and H is the Heaviside step function. In this study we wish to compare the performance of ESP forecasts with $m = 25$ ensemble members to FoGSS forecasts with $M = 1000$ ensemble members. When the CDF of the forecast is estimated empirically from an ensemble, CRPS calculations are sensitive to ensemble size, with greater errors occurring in smaller ensembles (e.g. Zamo and Naveau, 2018). Ferro et al. (2008) derived an unbiased estimator for CRPS for cases where ensemble members are exchangeable:

$$C_m = \frac{M(m+1)}{m(M+1)} C_M \quad (21)$$

where C_M is calculated on ensembles of $M = 1000$ with Eq (20), and C_m is an estimate for the smaller ensemble size $m = 25$.

CRPS is commonly presented as a skill score, where forecast accuracy in a forecasting system is calculated against a reference forecast:

$$CRPSS = 1 - \frac{C_{\mathrm{m,Fct}}}{C_{\mathrm{m,Ref}}} \quad (22)$$

CRPSS ranges from - ∞ (worst performance) to ∞ (best performance), with values near zero indicating the forecast performs similarly to the reference. We use climatology forecasts as our reference. Climatology forecasts are generated by randomly drawing 1000 realisations from a log-sinh transformed normal distribution fitted to each calendar month. We use the

BJP to fit the log-sinh transformation, following the data and cross-validation scheme described in sections 5 and 6.1. The BJP generates reliable distributions even where many zeros are present (Wang and Robertson, 2011). We also assess the skill of volume forecasts accumulated over multiple months. To calculate the climatology reference forecasts in these cases, we first accumulate volumes from the observed record, and then generate a climatology as described above.

To assess the significance of skill, we bootstrap equations (20)-(22) with 500 repeats. When 97.5% of bootstrapped CRPSS values are positive (negative), we consider the forecasts to be significantly skilful (negatively skilful).

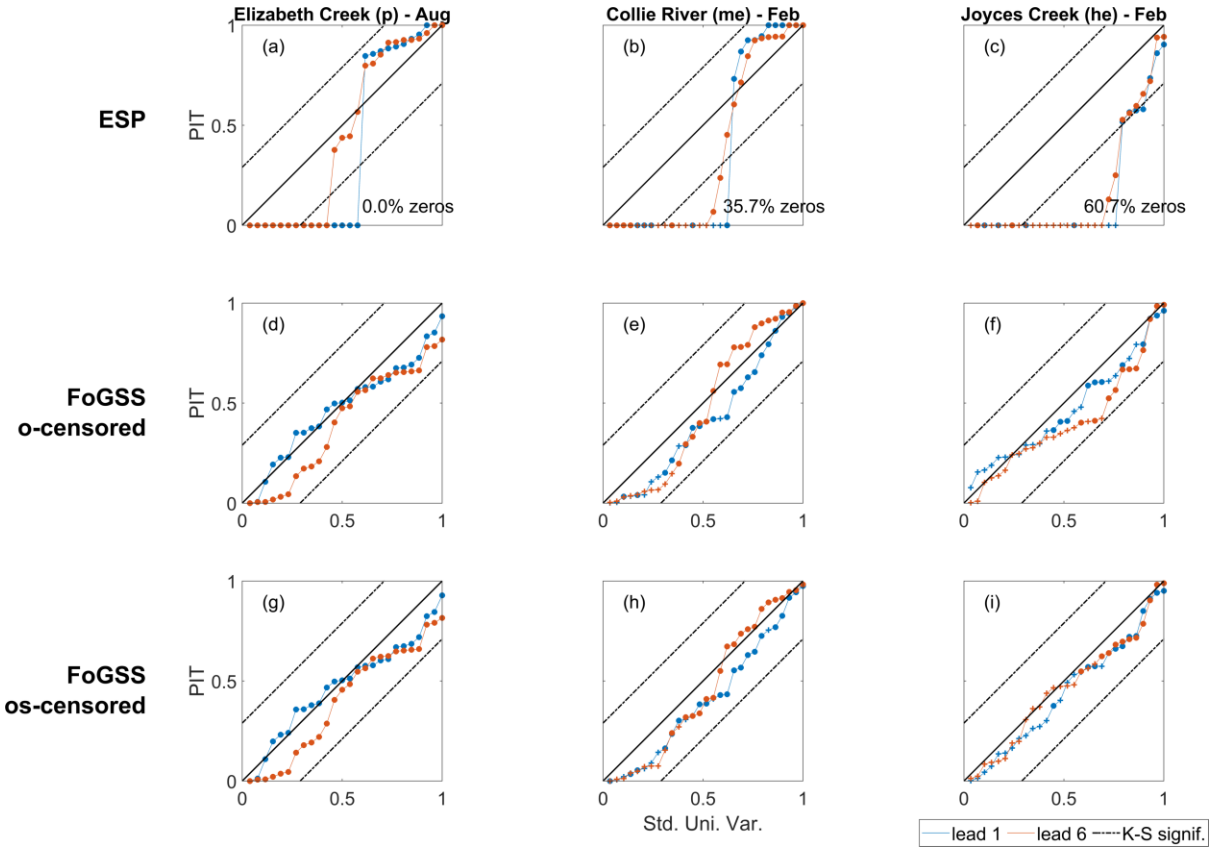


Figure 4 PIT uniform probability plots for forecasts for a dry month in example catchments (columns). Insets in (a)-(c) show proportion of zero flows in that month. Points show PIT values, crosses show pseudo-PIT values. Dashed lines give 95% Kolmogorov-Smirnoff confidence intervals. Top row shows ESP forecasts, middle row shows forecasts generated with the previous FoGSS method (o-censored) and bottom row shows new FoGSS method (os-censored).

7 Results

7.1 Reliability

Reliability of ESP and FoGSS forecasts for a dry month is shown in Fig 4 for our three example catchments (all 50 catchments are presented in Fig S1). Perfectly reliable forecasts follow the 1-1 line in the PIT diagrams. As foreshadowed in the introduction, ESP forecasts are often overconfident (the PIT diagram is s-shaped) because they do not consider uncertainties in the conversion of rainfall to runoff (e.g., uncertainties in model states, structure and/or streamflow observations). Overconfidence is particularly prevalent for the dry months when rainfall is low and streamflow results primarily from catchment stores draining rainfall accrued in previous (often wetter) months. This means forcing uncertainty

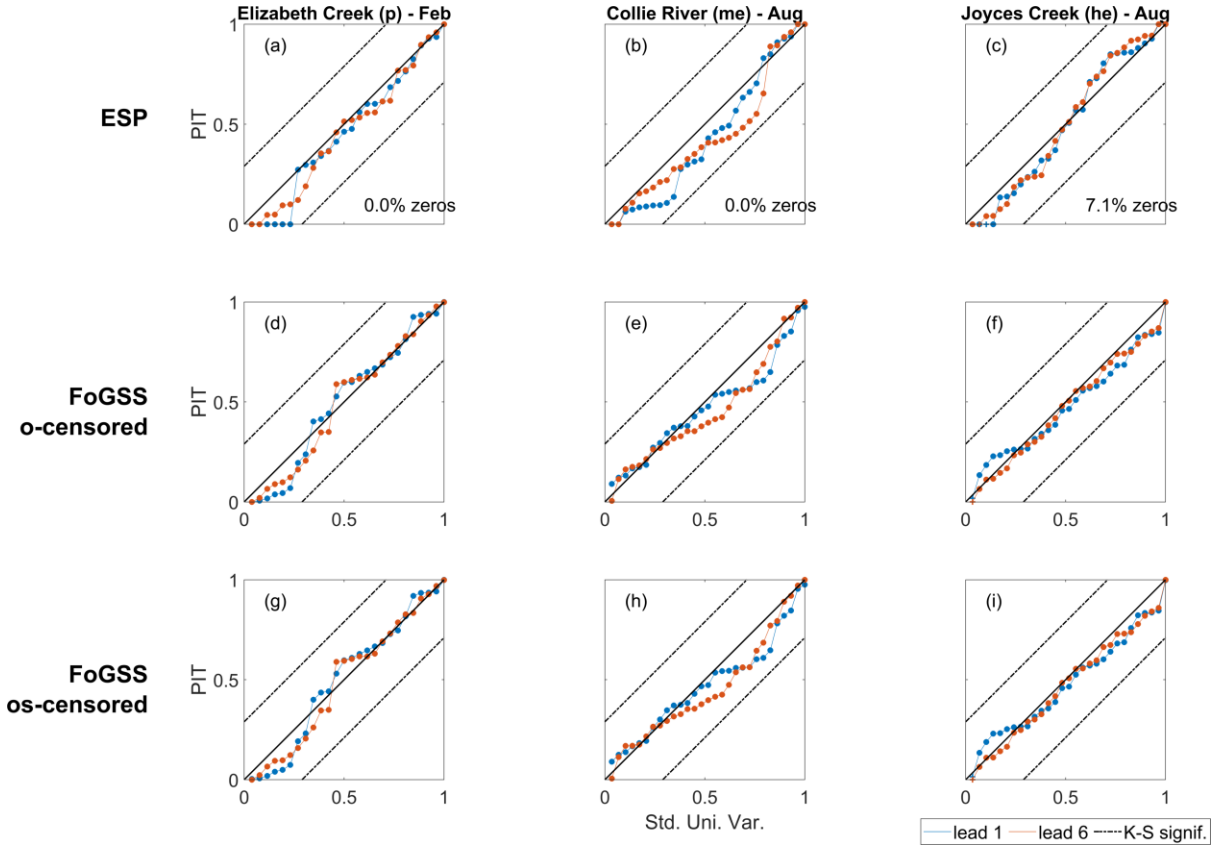


Figure 5 PIT uniform probability plots for forecasts for a wet month in example catchments (columns). Insets in (a)-(c) show proportion of zero flows in that month. Points show PIT values, crosses show pseudo-PIT values. Dashed lines give 95% Kolmogorov-Smirnoff confidence intervals. Top row shows ESP forecasts, middle row shows forecasts generated with the previous FoGSS method (o-censored) and bottom row shows new FoGSS method (os-censored).

has little bearing on overall uncertainty. At shorter lead times, forecast uncertainty in dry months is dominated by uncertainty in initial hydrological conditions. ESP reliability is considerably better for wetter months (Fig 5) than for drier months. In wetter months rainfall forcings are often a dominant source of uncertainty, and thus ESP ensembles do a reasonably good job of representing total uncertainty.

Both previous and new FoGSS methods (o-censoring and os-censoring, respectively) correct overconfidence in the ESP ensembles in many cases, producing reliable forecasts at short and long lead times in most instances. Improvements are most striking in dry months (Fig 4), where poor reliability in ESP is markedly improved by FoGSS. Even in wetter months, however, we can see small improvements in reliability after FoGSS is applied. At lead 1 in the perennial Elizabeth Creek catchment (Fig 5a), six observations are below the ESP forecast ensemble (these points fall on the bottom of the vertical axis). While these are technically within our confidence intervals, to have six forecasts (of 29) that completely miss an observation is evidence of a poorly performing forecast. FoGSS completely corrects this problem (Fig 5d).

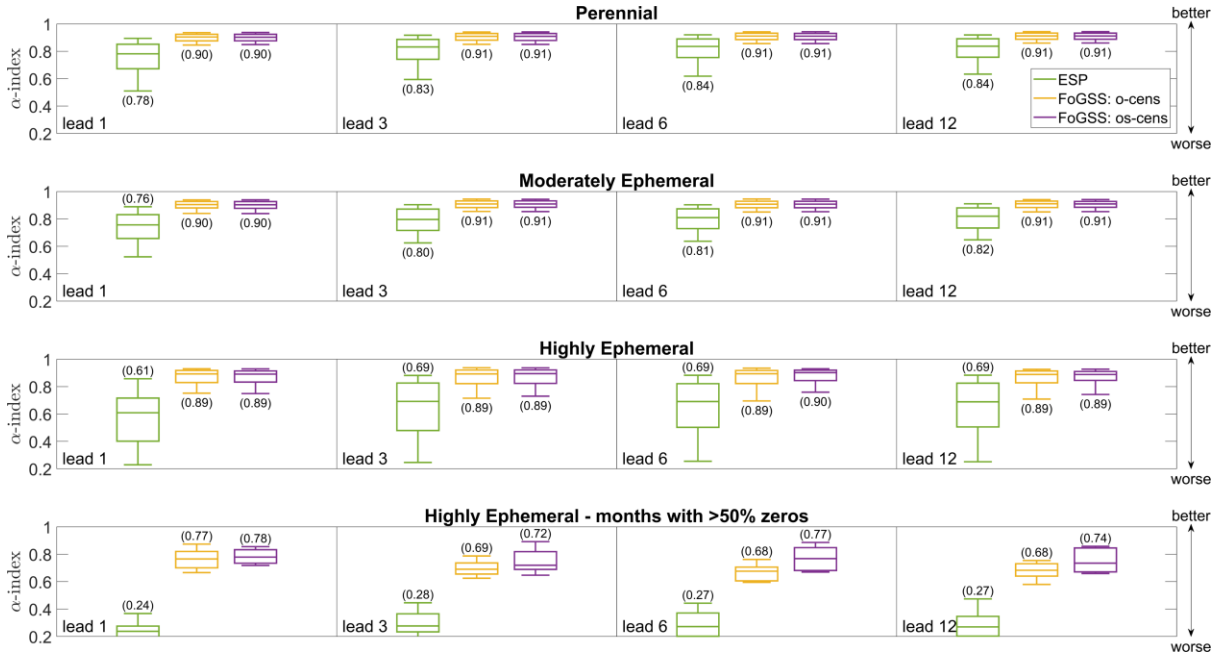


Figure 6 Alpha index with lead time summarised in box plots for all month in perennial (top), moderately ephemeral (upper middle) and highly ephemeral (lower middle) catchments. Bottom panel shows performance only for months with >50% zeros. Columns show lead times. Boxes give interquartile range and median, whiskers show 90% range. Numbers in brackets give median value for the corresponding box.

The benefits of os-censoring over o-censoring are apparent at longer lead times in dry months in the highly ephemeral Joyce catchment (Fig 4f,i). FoGSS with os-censoring produces reliable ensemble forecasts in the highly ephemeral Joyce River catchment in a month with >60% zeros (Fig 4f), regardless of lead time. For o-censoring, at lead 1 the forecasts are slightly underconfident (ensembles are too wide), signified by the transposed s-shape of the PIT diagram. At lead 6 the forecasts are increasingly positively biased; a result of accumulated instances of underconfident forecasts at multiple lead times. We note that o-censoring still performs reasonably well in this catchment, but this is less true of other catchments, as discussed in the next paragraph.

The improved performance of os-censoring over o-censoring becomes more evident in highly ephemeral months when we summarise reliability at a range of lead times (Fig 6). Both o-censored and os-censored versions of FoGSS markedly outperform ESP at short-lead times, especially in highly ephemeral rivers. ESP forecasts tend to become more reliable (closer to the ideal value of 1) with lead time because uncertainties from the forcings explain a large proportion of total uncertainty at long lead times. The benefits of os-censoring over o-censoring are marked in months with >50% zeros at longer lead times (Fig 6, bottom panel). Os-censoring allows FoGSS to maintain reliability in even highly ephemeral months to long lead times, whereas the reliability of o-censored FoGSS forecasts drops with lead time, as instances of underconfidence/positive bias accumulate. Consistent with Wang et al. (2020), we do not expect (and do not see) strong differences in skill or the reliability of accumulated volumes between o-censored and os-censored FoGSS forecasts; we therefore concentrate on os-censored forecasts in the remaining figures.

As stated in the introduction, a key advantage of forecasts in the form of time series is that they can be summed to produce total volume forecasts. Reliability of total volume forecasts is not ensured by reliability at individual lead times; it can only be guaranteed if hydrographs in the ensemble have realistic temporal properties (Demargne et al., 2014). Fig 7 shows that ESP forecasts can achieve reasonably reliable 12-month aggregations, but reliability falls away for shorter aggregation periods, irrespective of ephemerality. FoGSS forecasts of aggregated values are strongly reliable, and reliability is consistent across different aggregation periods (and across different catchments). Reliability of aggregated volumes in highly ephemeral catchments is slightly less than that of perennial and moderately ephemeral rivers, but is still high ($\alpha > 0.8$ in a large majority of cases, where $\alpha = 1$ is perfect reliability).

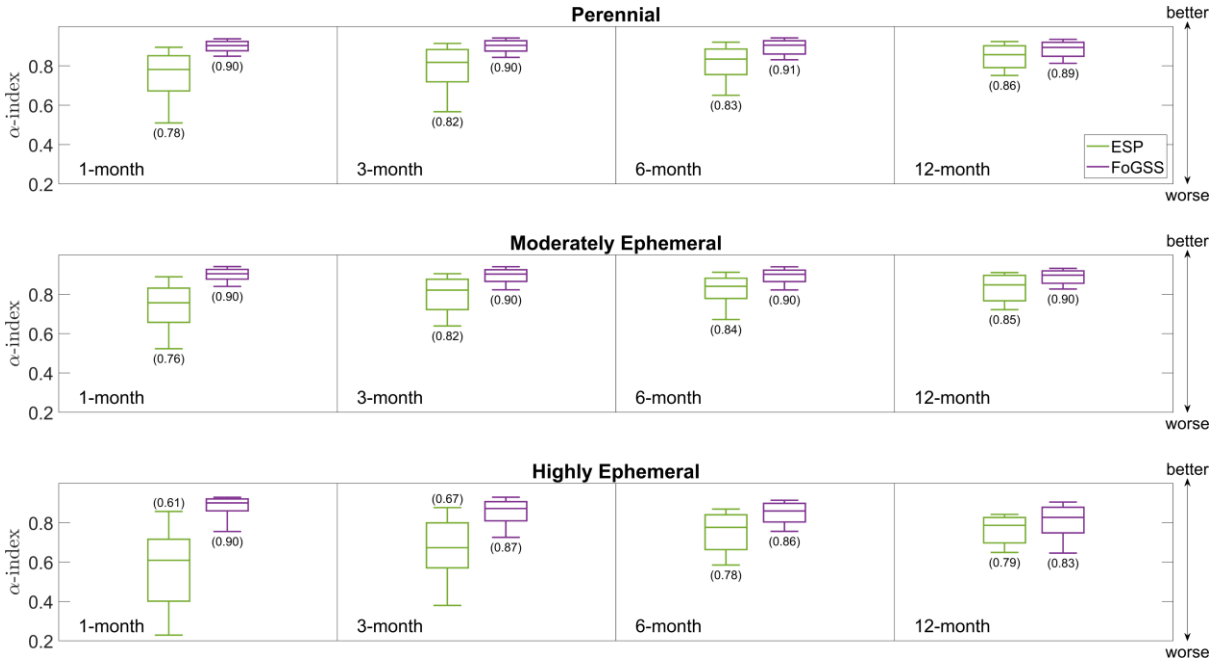
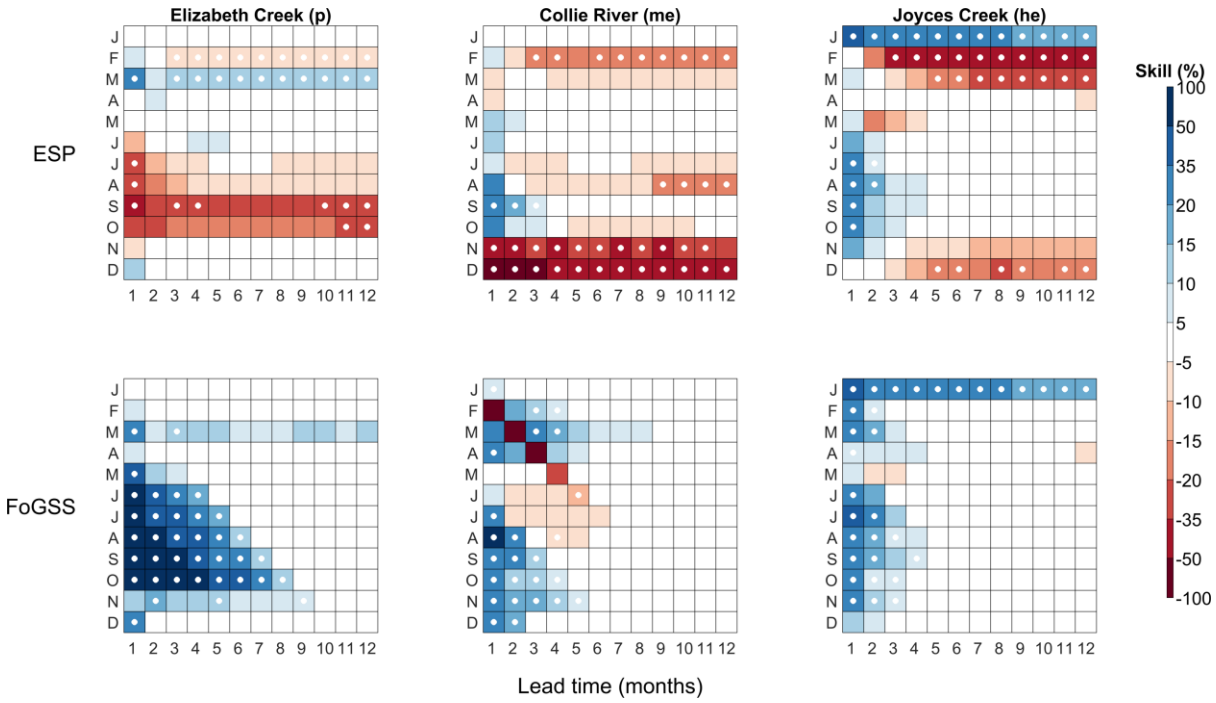


Figure 7 Alpha index of lead 1 forecasts of aggregated volumes. Alpha indices are summarised in box plots for forecasts in perennial (top), moderately ephemeral (middle) and highly ephemeral (bottom) catchments. Columns show aggregated volumes over different periods. Colours show ESP (green) and os-censored FoGSS (purple) forecasts. Boxes give interquartile range and median, whiskers show 90% range. Numbers in brackets give median value for the corresponding box.

7.2 Accuracy and skill

While ESP forecasts can be skilful to multiple lead times, they are beset by statistically significant negative skill in many months and lead times, irrespective of catchment ephemerality (Fig 8). Significant negative skills are present in Jul-Oct and Feb in Elizabeth Creek; Aug, Nov, Dec and Feb in the Collie River; and Dec, Feb and Mar in Joyces Creek. FoGSS forecasts, by contrast, virtually always produce positively skillful forecasts at short lead times and neutrally skilful forecasts (i.e., similarly skilful to climatology) at longer lead times – the one exception in June at lead 5 in the Collie catchment, where os-censored FoGSS produces significant negative skill but the ESP forecasts do not. This is consistent with the findings of our previous work in perennial and moderately ephemeral catchments (Bennett et al., 2016b; Bennett et al., 2017). It confirms that stages 2 and 3 of FoGSS substantially improve biases and reduce errors at short lead times, respectively.



680

681 Figure 8 Continuous ranked probability skill scores of target months for three example
682 catchments. Ephemerality is denoted by p-perennial (left column), me-moderately ephemeral
683 (middle column) and he-highly ephemeral (right column). White points indicating statistical
684 significance at the 5% level under bootstrapping (see text for details). Top row shows ESP;
685 bottom row shows os-censored FoGSS. See text for discussion of prominent features.

686

687 Fig 8 also shows that os-censored FoGSS forecasts for the highly ephemeral Joyces Creek
688 catchment are now positively or neutrally skilful in all months. This is also true of other
689 highly ephemeral catchments: FoGSS improves skill of ESP forecasts for >85% of highly
690 ephemeral months used in this study (Fig S4). Indeed FoGSS almost always improves ESP
691 forecasts, often substantially, with few exceptions (Fig 8, Figs S3 & S4). For example,
692 negative skills in the perennial Elizabeth Creek in Aug-Oct are completely removed (Fig 8),
693 and earlier lead times for these months are now significantly skilful. Similar improvements
694 are evident for Nov-Dec in the moderately ephemeral Collie River and for Feb-Mar in the
695 highly ephemeral Joyce Creek.

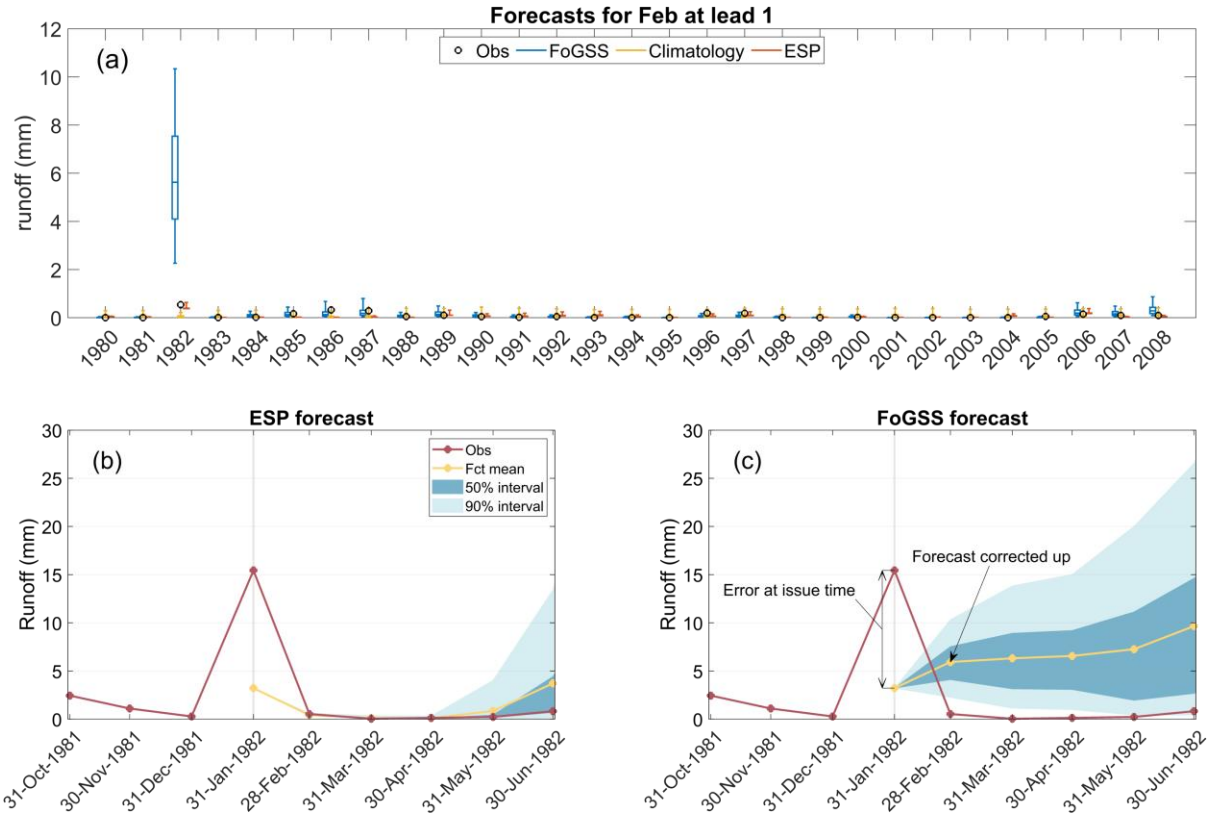


Figure 9 Cause of negative skill scores for Lead 1 forecasts for February in the Collie River. (a) Shows that negative forecast skill is caused by a single poor forecast in 1982; (b) shows ESP forecast issued for Feb 1982 and (c) shows os-censored FoGSS forecast issued for Feb 1982, and how autoregressive updating causes this poor forecast (see text for details).

Some positive skills are caused by poor performance of the reference forecast. A notable example is the significant positive skill of both ESP and FoGSS forecasts for January for Joyces Creek, even to very long lead times. In this case, the BJP produces a very wide climatology ensemble for January. This is caused by one very large event in the streamflow record (in what is usually a dry month), which affects the fit of the log-sinh transformation by the BJP. FoGSS is less affected, because in FoGSS we fit the transformation to data from all months. The BJP's poor performance in this case is highly unusual: more often, fitting climatology distributions by month leads to more accurate forecasts. Further, fitting the transformation to all months in FoGSS sometimes leads to a poorly fitting transformation at individual months. In exceptional cases this can result in negative skill in FoGSS forecasts (e.g. forecasts for Nov in the 922001A gauge, Fig S4).

There is one notable exception to FoGSS improving ESP forecasts in the Collie River: strongly negative skill has been introduced by FoGSS at lead 1 in Feb and follows a diagonal

to Lead 4 in May. These negative skills are not statistically significant, but demonstrating their cause is instructive. Fig 9 shows how negative forecast skill is introduced by FoGSS for the Collie River. Negative skill for the lead 1 Feb forecasts is caused by a single large overestimation of streamflow in 1982 (Fig 9a), and this also causes negative skill at lead 2 Mar forecasts through to lead 4 May forecasts. The large overestimation in Feb 1982 is due to the underestimation of a very large and unseasonal streamflow event in Jan 1982 driven by Tropical Cyclone Bruno (http://www.bom.gov.au/cyclone/history/bruno.shtml). Fig 9b shows that this has virtually no impact on the ESP forecast for lead 1 Feb (or the subsequent lead times). The AR1 model in FoGSS, however, corrects forecasts for Feb 1982 up dramatically (Fig 9c), following the mismatch in simulations and observations in Jan 1982. Under strict cross-validation, poor performance of a calibrated hydrological model can be difficult to avoid for isolated extreme events. In addition, the AR1 model in FoGSS exacerbates the poor performance of the hydrological model in this instance. However, while the AR1 model exacerbates poor model performance in this instance (as well as a few other isolated instances in Fig S4), the AR1 model overwhelmingly improves forecast accuracy in the vast majority of cases (e.g. in Elizabeth Creek for May-Oct).

Figure 10 Continuous Ranked Probability Skill scores with lead time summarised in box plots for all months in perennial (top), moderately ephemeral (middle) and highly ephemeral (bottom) catchments. Columns show lead times. Colours show ESP (green) and os-censored FoGSS (purple) forecasts. Boxes give interquartile range and median, whiskers show 90% range. Numbers in brackets give median value of each box.

FoGSS clearly improves the accuracy of ESP forecasts across the 50 catchments tested in this study (Fig 10). ESP forecasts for highly ephemeral rivers are noticeably less accurate than forecasts for perennial or moderately ephemeral rivers at all lead times. Conversely, FoGSS forecasts for highly ephemeral catchments tend to be similarly accurate as those for perennial and moderately ephemeral rivers. Crucially, at long lead times FoGSS forecasts tend to be neutrally skillful (i.e., similarly skillful to climatology), whereas ESP forecasts can be negatively skillful, in particular for highly ephemeral catchments.

Removing negative skills at long lead times has considerable benefits for the accuracy of aggregated volume forecasts (Fig 11). Skill at short lead times can carry through the entire volume forecast, resulting in FoGSS forecasts that can be skillful even for 12-month aggregations, including in highly ephemeral catchments. FoGSS forecasts for 6-month aggregations are very often skillful, irrespective of ephemerality. As with individual lead times, FoGSS forecasts of aggregated volumes are almost always more skilful than ESP forecasts of aggregated volumes.

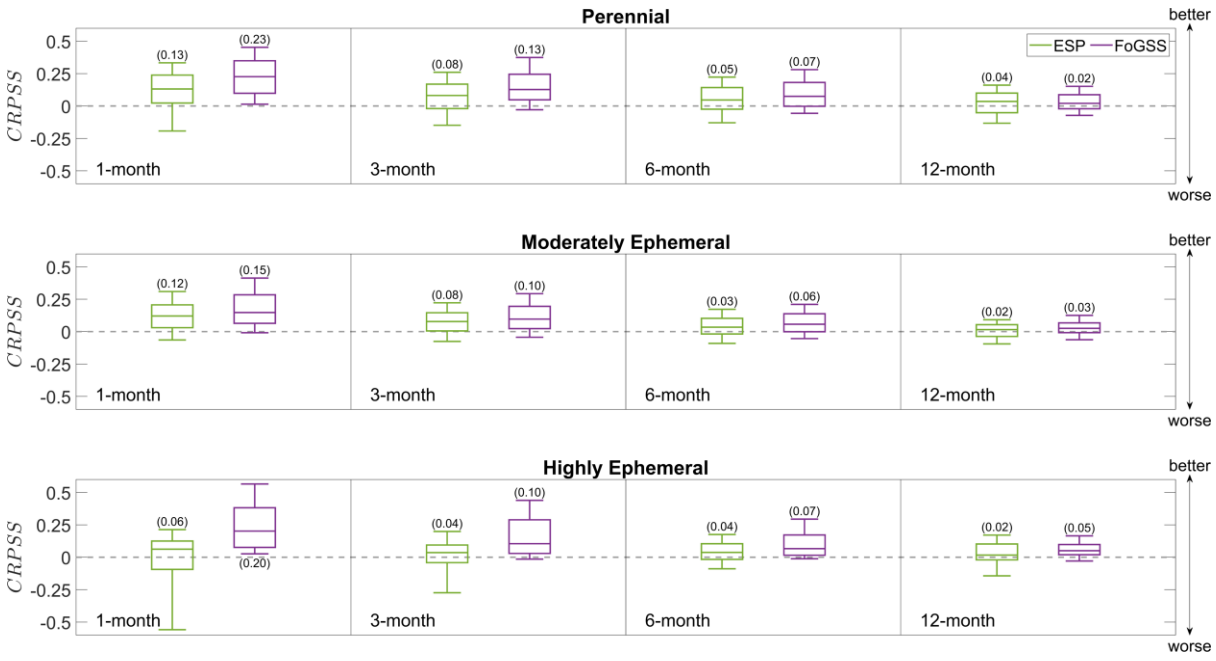


Figure 11 Continuous Ranked Probability Skill scores of lead 1 forecasts of aggregated volumes summarised for all months for perennial (top), moderately ephemeral (middle) and highly ephemeral (bottom) catchments. Columns show aggregation periods. Colours show ESP (green) and os-censored FoGSS (purple) forecasts. Boxes give interquartile range and median, whiskers show 90% range. Numbers in brackets give median value for the corresponding box.

8 Discussion

With the addition of the data censoring treatment described in this study, FoGSS now produces 12-month streamflow forecasts at the monthly time step that are reliable in highly ephemeral rivers. Forecasts are in the form of an ensemble of time series, where each ensemble member can be aggregated to produce ensemble forecasts of aggregated volumes. Forecasts are reliable at individual lead times and for aggregated volumes. As with previous versions of FoGSS, forecasts are almost always more skilful than ESP forecasts. Crucially, as skill declines with lead time, forecasts become neutrally skillful. This allows aggregated volume forecasts to be skillful, even to very long (e.g. 6-month) aggregation periods. These properties remove many of the barriers to the use of long-range ensemble streamflow forecasts in ephemeral rivers. We have shown in other work that FoGSS forecasts can improve the management of dams (Turner et al., 2017) and the allocation of water (Kaune et al., 2020), and these benefits can now be realised for highly ephemeral rivers.

This study reaffirms that it is possible to use FoGSS error model parameters estimated at lead 1 – analogous to a hydrological model calibration – to calibrate forecasts to multiple lead times. Crucially, this allows forecasts to be represented as a hydrograph, rather than a set of discrete probability distributions at each lead time. This is possible through the use of stochastic updating. Stochastic updating assumes that for a given month, the autocorrelation properties and error distributions at lead 1 hold at all lead times. This assumption is not guaranteed to hold, and accordingly FoGSS cannot enforce reliability at each month/lead time in the way that statistical methods can – although we have shown that in practice it almost always does. It is possible, then, that statistical forecasts that account for zero values (like the BJP; Wang and Robertson, 2011) will produce more reliable forecasts at discrete lead times (noting again that there is no clear method to generate hydrographs with statistical methods). In future work, we plan to compare climate-forced os-censored FoGSS forecasts with statistical forecasts generated with the BJP, where we adapt BJP forecasting methods to produce monthly forecasts to 12 months.

We note that there are aspects of the FoGSS error model that may still be improved. For example, McInerney et al. (2019) found that they could achieve slightly sharper predictions using a fixed-parameter Box-Cox transformation instead of a log-sinh transformation in ephemeral rivers. Conversely, the log-sinh transformation performs strongly in perennial catchments in comparison to other transformations (McInerney et al., 2017; Wang et al.,

2012). It is therefore possible that a different transformation may improve the properties of the FoGSS ensemble in ephemeral rivers.

In this study we have used uninformative forcings to drive our streamflow forecasts. FoGSS completely separates uncertainties in hydrological modelling from uncertainties in forcings, so it is relatively straightforward to include informative forcings from seasonal climate forecasts. However, to ensure reliable streamflow forecasts, forcings must also be reliable. Only then can uncertainties from forcings and hydrological models sum to correctly represent total forecast uncertainty. In addition, forcings should be coherent (i.e., at least as skilful as climatology) to ensure that streamflow forecast skill does not become negative at long lead times. ESP forcings are inherently reliable and coherent, but this is not necessarily true of climate forecasts (e.g., Peng et al., 2014; Schepen et al., 2016; Strazzo et al., 2019; Wang et al., 2019). Thus climate forecasts should only be used with FoGSS after they have been formally statistically calibrated, for which a number of methods are available (Manzanas et al., 2019; Sansom et al., 2016; Schepen and Wang, 2014; Siegert and Stephenson, 2019). Simple bias-corrections do not ensure reliability or coherence, and are not suitable for calibrating forecasts (Zhao et al., 2017).

In our previous applications of FoGSS, we have used calibrated climate forecasts as forcings, and forecasts were reliable in perennial and moderately ephemeral catchments (Bennett et al., 2016b; Bennett et al., 2017). However, these studies differed from the present one in that they used a hydrological model run at the monthly time step, not at the daily time step. Calibrating climate forecasts at a monthly time step is much more straightforward than at the daily time step, as monthly data are far less noisy. While it is technically possible to generate calibrated daily climate forecasts (Schepen et al., 2017), these methods have never been tested to the very long lead times (i.e. 365 days) used in our study. In addition, propagating uncertainty with stochastic updating – as used in FoGSS – is more difficult over many lead times (Bennett et al., 2021; Li et al., 2020) and is likely to require further development to produce forecasts for 365 lead times compared to the 12 lead times in this study. Testing and further development of daily climate forecast calibration methods and error modelling to very long lead times is thus a clear target for future research.

FoGSS is not yet in operational use, but offers an attractive means for a future operational long-range streamflow forecasting system. As we have shown, os-censoring works across all catchment types; in perennial catchments censoring is not enacted, while in ephemeral

catchments it plays an important role. This means that a single method can be applied to all catchment types – a clear advantage for operations. Further, we have shown that FoGSS is effective as a 'bolt-on' corrective to ESP forecasts – in this case for a hydrological model calibrated using an independent procedure. As ESP forecasts are widely used in operational forecasting, FoGSS can be easily applied to improve these forecasts. Correcting ESP forecasts is also attractive as ESP can be readily adapted to accept calibrated climate forecast inputs. In other words, because FoGSS is premised on the idea of separating the uncertainties in climate forecasts from uncertainties in hydrological modelling, this makes any FoGSS operational system highly modular: climate forecasts can be updated/improved independently of hydrological models, and vice versa. We note also that FoGSS is highly computationally efficient, often a key consideration for operationalisation: on a standard desktop computer, it takes <2 minutes to estimate parameters for the 29-year estimation period used in this study, while it takes at most a few seconds to generate a 12-month forecast with 1000 ensemble members.

9 Summary and Conclusions

This study further develops the FoGSS (forecast guided stochastic scenarios) method to generate long-range (12-month) streamflow forecasts for use in highly ephemeral rivers. Forecasts are in the form of an ensemble of time series at the monthly time step. Each ensemble member can be summed to produce an ensemble of aggregated volume forecasts. We combine the basic staged structure of the FoGSS error model with a data censoring method that is applied to both observations and simulations. The data censoring method treats both modelled and simulated flow as censored data when they are equal to zero. This allows FoGSS to generate reliable ensemble forecasts in highly ephemeral rivers, which can cease to flow >50% of the time in some months.

We test the new version of FoGSS on 50 catchments, of which 26 are ephemeral. FoGSS forecasts are generally highly skillful at short lead times compared to climatology, and very often more skillful than conventional ESP (Ensemble Streamflow Prediction) forecasts. At longer lead times FoGSS forecasts are 'coherent' – that is, never less skillful than climatology – which cannot be guaranteed by ESP forecasts. Forecasts are reliable irrespective if they are issued for highly ephemeral (>50% zeros) or perennial (<5% zeros) months. Forecasts perform similarly in highly ephemeral, moderately ephemeral and perennial catchments in both skill and reliability. These improvements pave the way for

operational long-range forecasts in ephemeral rivers, meeting a key need for improved water management.

Acknowledgements

This work was conducted on the traditional lands of the Boonwurrung and Wurundjeri peoples of the Kulin Nation. We acknowledge their continuing custodianship of these lands and the rivers that flow through them, and pay our respects to their elders, past and present. We also acknowledge the traditional custodians of the catchments and rivers used in this study.

This research was supported by the Water Information Research And Development Alliance (WIRADA) between the Bureau of Meteorology and CSIRO Land & Water, and ARC linkage project LP170100922. Thanks to Elisabeth Vogel and Richard Laugesen (both Bureau of Meteorology) for helpful comments on the manuscript.

All data and model runs used in this paper are available from CSIRO's data access portal at http://dx.doi.org/[DOI_NO_STILL_TO_COME]. Matlab and C++ code used to generate simulations and forecasts, and to verify forecasts, are available on request; license conditions apply.

References

Ammann, L., Fenicia, F., Reichert, P., 2019. A likelihood framework for deterministic hydrological models and the importance of non-stationary autocorrelation. Hydrol. Earth Syst. Sci., 23(4): 2147-2172. DOI:10.5194/hess-23-2147-2019

Arnal, L. et al., 2018. Skilful seasonal forecasts of streamflow over Europe? Hydrol. Earth Syst. Sci., 22(4): 2057-2072. DOI:10.5194/hess-22-2057-2018

Bennett, J.C., Robertson, D.E., Wang, Q.J., Li, M., Perraud, J.-M., 2021. Propagating reliable estimates of hydrological forecast uncertainty to many lead times. Submitted to Journal of Hydrology.

Bennett, J.C., Robertson, D.E., Ward, P.G.D., Hapuarachchi, H.A.P., Wang, Q.J., 2016a. Calibrating hourly rainfall-runoff models with daily forcings for streamflow forecasting applications in meso-scale catchments. Environmental Modelling & Software, 76: 20-36. DOI:10.1016/j.envsoft.2015.11.006

Bennett, J.C., Wang, Q.J., Li, M., Robertson, D.E., Schepen, A., 2016b. Reliable long-range ensemble streamflow forecasts: Combining calibrated climate forecasts with a conceptual runoff model and a staged error model. Water Resources Research, 52: 8238–8259. DOI:10.1002/2016wr019193

Bennett, J.C. et al., 2017. Assessment of an ensemble seasonal streamflow forecasting system for Australia. Hydrol. Earth Syst. Sci., 21(12): 6007-6030. DOI:10.5194/hess-21-6007-2017

Chatterjee, S., McLeish, D.L., 1986. Fitting linear regression models to censored data by least squares and maximum likelihood methods. Communications in Statistics - Theory and Methods, 15(11): 3227-3243. DOI:10.1080/03610928608829305

Coron, L. et al., 2012. Crash testing hydrological models in contrasted climate conditions: An experiment on 216 Australian catchments. Water Resources Research, 48(5): W05552. DOI:10.1029/2011wr011721

Costigan, K.H. et al., 2017. Chapter 2.2 - Flow Regimes in Intermittent Rivers and Ephemeral Streams. In: Datry, T., Bonada, N., Boulton, A. (Eds.), Intermittent Rivers and Ephemeral Streams. Academic Press, pp. 51-78. DOI:10.1016/B978-0-12-803835-2.00003-6

Crochemore, L., Ramos, M.H., Pappenberger, F., 2016. Bias correcting precipitation forecasts to improve the skill of seasonal streamflow forecasts. Hydrology and Earth System Sciences, 20: 3601-3618. DOI:10.5194/hess-2016-78

Datry, T., Bonada, N., Boulton, A.J., 2017. Chapter 1 - General Introduction. In: Datry, T., Bonada, N., Boulton, A. (Eds.), Intermittent Rivers and Ephemeral Streams. Academic Press, pp. 1-20. DOI:10.1016/B978-0-12-803835-2.00001-2

Day, G.N., 1985. Extended streamflow forecasting using NWSRFS. Journal of Water Resources Planning and Management, 111(2): 157–170. DOI:10.1061/(ASCE)0733-9496(1985)111:2(157)

Demargne, J. et al., 2014. The Science of NOAA's Operational Hydrologic Ensemble Forecast Service. Bulletin of the American Meteorological Society, 95: 79–98. DOI:10.1175/bams-d-12-00081.1

Duan, Q.Y., Gupta, V.K., Sorooshian, S., 1993. Shuffled complex evolution approach for effective and efficient global minimization. Journal of Optimization Theory and Applications, 76(3).

Feikema, P.M. et al., 2018. Service and Research on Seasonal Streamflow Forecasting in Australia. In: Jung, H.-S., Wang, B. (Eds.), Bridging Science and Policy Implication for Managing Climate Extremes. World Scientific Series on Asia-Pacific Weather and Climate. World Scientific, pp. 157-175. DOI:doi:10.1142/9789813235663_0010

10.1142/9789813235663_0010
Ferro, C.A.T., Richardson, D.S., Weigel, A.P., 2008. On the effect of ensemble size on the discrete and continuous ranked probability scores. Meteorological Applications, 15(1): 19-24. DOI:10.1002/met.45
Gneiting, T., Katzfuss, M., 2014. Probabilistic forecasting. Annual Review of Statistics and Its Application, 1(1): 125-151. DOI:10.1146/annurev-statistics-062713-085831
Grafton, R.Q., Wheeler, S.A., 2018. Economics of Water Recovery in the Murray-Darling Basin, Australia. Annual Review of Resource Economics, 10(1): 487-510. DOI:10.1146/annurev-resource-100517-023039
Hemri, S., Klein, B., 2017. Analog-Based Postprocessing of Navigation-Related Hydrological Ensemble Forecasts. Water Resources Research, 53(11): 9059-9077. DOI:10.1002/2017wr020684
Jones, D.A., Wang, W., Fawcett, R., 2009. High-quality spatial climate data-sets for Australia. Australian Meteorological and Oceanographic Journal, 58: 233-248.
Kaune, A., Chowdhury, F., Werner, M., Bennett, J., 2020. The benefit of using an ensemble of seasonal streamflow forecasts in water allocation decisions. Hydrol. Earth Syst. Sci. Discuss., 2020: 1-32. DOI:10.5194/hess-2020-60
Krzysztofowicz, R., 1999. Bayesian theory of probabilistic forecasting via deterministic hydrologic model. Water Resources Research, 35(9): 2739-2750. DOI:doi:10.1029/1999WR900099
Li, M., Robertson, D.E., Wang, Q.J., Bennett, J.C., Perraud, J.-M., 2020. Reliable hourly streamflow forecasting with emphasis on ephemeral rivers. Journal of Hydrology, 125739. DOI:10.1016/j.jhydrol.2020.125739
Li, M., Wang, Q.J., Bennett, J., 2013. Accounting for seasonal dependence in hydrological model errors and prediction uncertainty. Water Resources Research, 49: 5913–5929. DOI:10.1002/wrcr.20445
Li, M., Wang, Q.J., Bennett, J.C., Robertson, D.E., 2015. A strategy to overcome adverse effects of autoregressive updating of streamflow forecasts. Hydrology and Earth System Sciences, 19(1): 1-15. DOI:10.5194/hess-19-1-2015
Li, M., Wang, Q.J., Bennett, J.C., Robertson, D.E., 2016. Error reduction and representation in stages (ERRIS) in hydrological modelling for ensemble streamflow forecasting. Hydrology and Earth System Sciences, 20(9): 3561-3579. DOI:10.5194/hess-20-3561-2016
Li, M., Wang, Q.J., Robertson, D.E., Bennett, J.C., 2017. Improved error modelling for streamflow forecasting at hourly time steps by splitting hydrographs into rising and falling limbs. Journal of Hydrology, 555: 586-599. DOI:10.1016/j.jhydrol.2017.10.057
Liu, L., Wang, Q.J., Xu, Y.-P., 2020. Temporally varied error modelling for improving simulations and quantifying uncertainty. Journal of Hydrology, 586: 124914. DOI:10.1016/j.jhydrol.2020.124914
Manzanas, R. et al., 2019. Bias adjustment and ensemble recalibration methods for seasonal forecasting: a comprehensive intercomparison using the C3S dataset. Climate Dynamics, 53(3): 1287-1305. DOI:10.1007/s00382-019-04640-4
McInerney, D., Kavetski, D., Thyer, M., Lerat, J., Kuczera, G., 2019. Benefits of explicit treatment of zero flows in probabilistic hydrological modelling of ephemeral catchments. Water Resources Research, 0(ja). DOI:10.1029/2018WR024148
McInerney, D., Thyer, M., Kavetski, D., Lerat, J., Kuczera, G., 2017. Improving probabilistic prediction of daily streamflow by identifying Pareto optimal approaches for modeling heteroscedastic residual errors. Water Resources Research, 53(3): 2199-2239. DOI:10.1002/2016wr019168

969 Pagano, T.C., Garen, D.C., Perkins, T.R., Pasteris, P.A., 2009. Daily Updating of Operational
970 Statistical Seasonal Water Supply Forecasts for the western U.S.1. JAWRA Journal of
971 the American Water Resources Association, 45(3): 767-778. DOI:10.1111/j.1752-
972 1688.2009.00321.x
973 Peng, Z. et al., 2014. Statistical calibration and bridging of ECMWF System4 outputs for
974 forecasting seasonal precipitation over China. Journal of Geophysical Research
975 (Atmospheres), 119: 7116–7135. DOI:10.1002/2013JD021162.
976 Perrin, C., Michel, C., Andréassian, V., 2003. Improvement of a parsimonious model for
977 streamflow simulation. Journal of Hydrology, 279: 275-289. DOI:10.1016/S0022-
978 1694(03)00225-7
979 Petheram, C., Gallant, J., Stone, P., Wilson, P., Read, A., 2018. Rapid assessment of potential
980 for development of large dams and irrigation across continental areas: application to
981 northern Australia. The Rangeland Journal, 40(4): 431-449. DOI:10.1071/RJ18012
982 Pokhrel, P., Robertson, D.E., Wang, Q.J., 2013. A Bayesian joint probability post-processor
983 for reducing errors and quantifying uncertainty in monthly streamflow predictions.
984 Hydrology and Earth System Sciences, 17(2): 795-804. DOI:10.5194/hess-17-795-
985 2013
986 Renard, B., Kavetski, D., Kuczera, G., Thyer, M., Franks, S.W., 2010. Understanding
987 predictive uncertainty in hydrologic modeling: The challenge of identifying input and
988 structural errors. Water Resources Research, 46(5): W05521.
989 DOI:10.1029/2009wr008328
990 Sansom, P.G., Ferro, C.A.T., Stephenson, D.B., Goddard, L., Mason, S.J., 2016. Best
991 Practices for Postprocessing Ensemble Climate Forecasts. Part I: Selecting
992 Appropriate Recalibration Methods. Journal of Climate, 29(20): 7247-7264.
993 DOI:10.1175/jcli-d-15-0868.1
994 Schepen, A., Wang, Q.J., 2014. 12 month out forecasts of catchment rainfall by post-
995 processing ECMWF System 4 and POAMA M2.4 forecasts, CSIRO Land and Water
996 for a Healthy Country Flagship, Highett.
997 Schepen, A., Wang, Q.J., Everingham, Y., 2016. Calibration, bridging and merging to
998 improve GCM seasonal temperature forecasts in Australia. Monthly Weather Review,
999 144(6): 2421-2441. DOI:10.1175/MWR-D-15-0384.1
1000 Schepen, A., Zhao, T., Wang, Q.J., Robertson, D.E., 2017. A new method for post-processing
1001 daily sub-seasonal to seasonal rainfall forecasts from GCMs and evaluation for 12
1002 Australian catchments. Submitted to Hydrology and Earth System Sciences.
1003 Seo, D.J., Herr, H.D., Schaake, J.C., 2006. A statistical post-processor for accounting of
1004 hydrologic uncertainty in short-range ensemble streamflow prediction. Hydrology and
1005 Earth System Sciences Discussions, 3: 1987-2035. DOI:10.5194/hessd-3-1987-2006
1006 Shrestha, D.L., Robertson, D.E., Bennett, J.C., Wang, Q.J., 2015. Improving precipitation
1007 forecasts by generating ensembles through postprocessing. Monthly Weather Review,
1008 143: 3642-3663. DOI:10.1175/mwr-d-14-00329.1
1009 Siegert, S., Stephenson, D.B., 2019. Chapter 15 - Forecast Recalibration and Multimodel
1010 Combination. In: Robertson, A.W., Vitart, F. (Eds.), Sub-Seasonal to Seasonal
1011 Prediction. Elsevier, pp. 321-336. DOI:https://doi.org/10.1016/B978-0-12-811714-
1012 9.00015-2
1013 Smith, T., Marshall, L., Sharma, A., 2015. Modeling residual hydrologic errors with Bayesian
1014 inference. Journal of Hydrology, 528: 29-37.
1015 DOI:https://doi.org/10.1016/j.jhydrol.2015.05.051
1016 Smith, T., Sharma, A., Marshall, L., Mehrotra, R., Sisson, S., 2010. Development of a formal
1017 likelihood function for improved Bayesian inference of ephemeral catchments. Water
1018 Resources Research, 46(12): W12551. DOI:10.1029/2010wr009514

Stern, H., Hoedt, G., Ernst, J., 2000. Objective classification of Australian climates. Australian Meteorological Magazine, 49: 87-96.
Strazzo, S. et al., 2019. Application of a Hybrid Statistical–Dynamical System to Seasonal Prediction of North American Temperature and Precipitation. Monthly Weather Review, 147(2): 607-625. DOI:10.1175/mwr-d-18-0156.1
Tooth, S., 2000. Process, form and change in dryland rivers: a review of recent research. Earth-Science Reviews, 51(1): 67-107. DOI:10.1016/S0012-8252(00)00014-3
Turner, S.W.D., Bennett, J.C., Robertson, D.E., Galelli, S., 2017. Complex relationship between seasonal streamflow forecast skill and value in reservoir operations. Hydrol. Earth Syst. Sci., 21(9): 4841-4859. DOI:10.5194/hess-21-4841-2017
Verkade, J.S., Brown, J.D., Davids, F., Reggiani, P., Weerts, A.H., 2017. Estimating predictive hydrological uncertainty by dressing deterministic and ensemble forecasts; a comparison, with application to Meuse and Rhine. Journal of Hydrology, 555(Supplement C): 257-277. DOI:10.1016/j.jhydrol.2017.10.024
Wang, Q., Shrestha, D.L., Robertson, D., Pokhrel, P., 2012. A log-sinh transformation for data normalization and variance stabilization. Water Resources Research, 48(5). DOI:http://dx.doi.org/10.1029/2011WR010973
Wang, Q.J., Bennett, J.C., Robertson, D.E., Li, M., 2020. A data censoring approach for predictive error modelling of flow in ephemeral rivers. Water Resources Research, 56: e2019WR026128. DOI:10.1029/2019WR026128
Wang, Q.J., Robertson, D.E., 2011. Multisite probabilistic forecasting of seasonal flows for streams with zero value occurrences. Water Resources Research, 47: W02546. DOI:10.1029/2010wr009333
Wang, Q.J., Robertson, D.E., Chiew, F.H.S., 2009. A Bayesian joint probability modeling approach for seasonal forecasting of streamflows at multiple sites. Water Resources Research, 45: W05407. DOI:10.1029/2008WR007355
Wang, Q.J. et al., 2019. An evaluation of ECMWF SEAS5 seasonal climate forecasts for Australia using a new forecast calibration algorithm. Environmental Modelling & Software, 122: 104550. DOI:https://doi.org/10.1016/j.envsoft.2019.104550
Woldemeskel, F. et al., 2018. Evaluating post-processing approaches for monthly and seasonal streamflow forecasts. Hydrol. Earth Syst. Sci., 22(12): 6257-6278. DOI:10.5194/hess-22-6257-2018
Wood, A.W., Schaake, J.C., 2008. Correcting Errors in Streamflow Forecast Ensemble Mean and Spread. J. Hydrometeorol., 9(1): 132-148. DOI:10.1175/2007JHM862.1
Yuan, X., Wood, E.F., Ma, Z., 2015. A review on climate-model-based seasonal hydrologic forecasting: physical understanding and system development. Wiley Interdisciplinary Reviews: Water, 2(5): 523-536. DOI:10.1002/wat2.1088
Zamo, M., Naveau, P., 2018. Estimation of the Continuous Ranked Probability Score with Limited Information and Applications to Ensemble Weather Forecasts. Mathematical Geosciences, 50(2): 209-234. DOI:10.1007/s11004-017-9709-7
Zhao, T. et al., 2017. How suitable is quantile mapping for post-processing GCM precipitation forecasts? Journal of Climate, 30(9): 3185-3196. DOI:10.1175/jcli-d-16-0652.1
Zhao, T., Schepen, A., Wang, Q.J., 2016. Ensemble forecasting of sub-seasonal to seasonal streamflow by a Bayesian joint probability modelling approach. Journal of Hydrology, 541, Part B: 839-849. DOI:http://dx.doi.org/10.1016/j.jhydrol.2016.07.040

Appendix A: Parameter estimation for each stage

Stage 1

Maximum a posteriori (MAP) estimation is used to estimate the transformation parameters in Eq (1). We assume that transformed observed streamflow follows a normal distribution

$$z_o \sim N\left(m, s^2\right) \tag{A1}$$

In Eq (1) and Eq (A1) a total of four parameters - a, b, m and s - are estimated. The four parameters are reparameterised to $\log(a)$, $\log(b)$, m/s and $\log(s)$, respectively, to make it more numerically efficient to find a MAP solution. Our parameter set is then $\theta_1 = \left\{\log(a), \log(b), m/s, \log(a)\right\}$. The likelihood is given by:

$$L(\theta_1) \propto p(\theta_1) \prod_t J_{z_o(t)\to cq_o(t)} N\left(z_o(t) \middle| m, s^2\right) \tag{A2}$$

where $J_{z_o(t)\to cq_o(t)}$ is the Jacobian

$$J_{z_o(t)\to cq_o(t)} = \coth\left(a + bcq_o(t)\right) \tag{A3}$$

with a standardization constant $c = 5/\max(\mathbf{q}_o)$, and $p(\theta_1)$ is a prior

$$p(\theta_1) \propto p\left(\log(a)\right) p(m/s)\, p\left(\log(s)\right) p\left(\log(b)\right), \tag{A4}$$

where

$$p\left(\log(a)\right) \propto 1, \quad \log(a) \le 0, \tag{A5}$$

$$p(m/s) \propto 1, \tag{A6}$$

$$p\left(\log(s)\right) \propto 1 \tag{A7}$$

and

$$p\left(\log(b)\right) \sim N\left(0, 1^2\right). \tag{A8}$$

The priors in equations (A5)-(A7) are uninformative. The prior on $\log(b)$ (Eq A8) encourages values closer to zero, which in our experience is a good starting point. If the datum at time t is a censored value (i.e., $q_o(t) = 0$), the term $J_{z_o(t)\to cq_o(t)} N\left(z_o(t) \middle| m, s^2\right)$ in Eq (A2) is replaced with

the normal cumulative probability $\Phi\left(z_c \middle| m, s^2\right)$, where $z_c = Tf(0)$ is the log-sinh transformed value of zero.

Stage 2

Parameters in Eq (2), $\theta_2 = \left\{d(i), \mu(i): \ i = 1,2,\ldots,12\right\}$, are estimated by minimizing the LST objective (Eq 7).

Stage 3

The $\theta_3 = \left\{\rho(i): \ i = 1,2,\ldots,12\right\}$ parameter in Eq (3) is estimated by minimizing the LST objective (Eq 7).

Stage 4

We first estimate $m_3(i)$ and $s_3^2(i)$. To do this, we generate a streamflow simulation $\mathbf{q}_3$ by applying stages 1-3 to the hydrological model simulation. We extract the simulated streamflow for each calendar month from $\mathbf{q}_3$, and then use the MAP procedure from Stage 1 to estimate $m_3(i)$, $s_3^2(i)$, keeping a and b fixed from the Stage 1 estimation.

We can then use maximum likelihood estimation to infer $\theta_4 = \left\{\sigma^2(i): \ i = 1,2,\ldots,12\right\}$ with the likelihood specified in Eq (12). Note that no closed form is available for $case = 4$ (Eq 11), and thus it requires a Monte Carlo integration. For details of this integration, see Wang et al. (2020).

Supplementary Materials

Table S1: Gauge site information

Gauge Name	Gauge Id	State	Area (km2)	Lon.	Lat.	Missing (%)	Ephemerality	No. months % zero >5	No. months % zero >50
Cockburn River at Mulla Crossing	419016	NSW	897	151.13	-31.06	0.9	Moderate	4	0
Flyers Creek at Beneree	412080	NSW	89	149.04	-33.51	20.5	Perennial	0	0
Goulburn River at Kerrabee	210016	NSW	4981	150.32	-32.42	14	Perennial	0	0
Gwydir River at Yarrowyck	418014	NSW	835	151.36	-30.47	0.9	Moderate	2	0
Jigadee Creek at Avondale	211008	NSW	64	151.47	-33.07	9.2	Moderate	2	0
Macintyre River at Inverell	416016	NSW	751	151.13	-29.79	0	Moderate	1	0
Maryland River downstream of Wylie Creek	204039	NSW	377	152.2	-28.47	4.8	Moderate	10	0
Maules Creek at Avoca East	419051	NSW	663	150.08	-30.5	0.9	Moderate	3	0
Mountain Creek at Mountain Creek	410107	NSW	185	147.85	-35.03	20.2	Moderate	6	0
Murray River at Biggara	401012	NSW	1165	148.05	-36.32	1.5	Perennial	0	0
Myrtle Creek at Rappville	203030	NSW	392	153	-29.11	0.3	Moderate	5	0
Nowendoc River at Rocks Crossing	208005	NSW	1870	152.08	-31.78	1.8	Perennial	0	0
Orara River at Karangi	204025	NSW	136	153.03	-30.25	4.2	Moderate	2	0
Paddy's River at Riverlea	410713	NSW	224	148.97	-35.38	2.1	Perennial	0	0
Shoalhaven River at Kadoona	215008	NSW	282	149.64	-35.79	8.3	Perennial	0	0
Styx River at Jeogla	206001	NSW	163	152.16	-30.59	3	Perennial	0	0
Towamba River at Towamba	220004	NSW	766	149.66	-37.07	1.2	Moderate	2	0
West Brook upstream of Glendon Brook	210080	NSW	72	151.28	-32.47	6.3	Moderate	6	0
Finniss River at Batchelor Dam Site	G8150010	NT	363	130.95	-13.02	0	Moderate	4	0

Archer River at Telegraph Crossing	922001A	QLD	2928	142.92	-13.42	14.3	High	5	1
Blencoe Creek at Blencoe Falls	116010A	QLD	224	145.54	-18.2	1.8	Perennial	0	0
Bremer River at Walloon	143107A	QLD	398	152.69	-27.6	10.7	Moderate	9	0
Daintree River at Bairds	108002A	QLD	907	145.28	-16.18	5.4	Perennial	0	0
Elizabeth Creek at Mount Surprise	917107A	QLD	528	144.31	-18.13	9.8	Perennial	0	0
Emu Creek at Emu Vale	422313B	QLD	153	152.23	-28.23	0.6	Moderate	1	0
Gregory River at Lower Gregory	122004A	QLD	46	148.55	-20.3	8.3	High	5	0
Mitchell River at Gamboola	919011A	QLD	20315	143.68	-16.54	8.9	Perennial	0	0
Tinana Creek at Tagigan Road	138009A	QLD	102	152.78	-26.08	6.3	Moderate	3	0
Walsh River at Rookwood	919310A	QLD	4927	144.29	-16.98	18.5	High	4	1
Waterpark Creek at Byfield	129001A	QLD	212	150.67	-22.84	5.4	Perennial	0	0
Brid River upstream of tidal limit	19200	TAS	138	147.37	-41.02	2.4	Perennial	0	0
Flowerdale River at Moorleah	14215	TAS	156	145.61	-40.97	10.7	Perennial	0	0
Scamander River upstream of Scamander water supply intake	2206	TAS	261	148.18	-41.45	20.8	Perennial	0	0
Gellibrand at Upper Gellibrand	235202	VIC	53	143.66	-38.56	0.3	Perennial	0	0
Holland Creek at Kelfeera	404207	VIC	448	146.06	-36.61	0	Moderate	4	0
Joyces Creek at Strathlea	407230	VIC	156	143.96	-37.16	0	High	11	3
Ovens River at Bright	403205	VIC	495	146.95	-36.73	0	Perennial	0	0
Tambo River at Swifts Creek	223202	VIC	943	147.72	-37.26	0	Perennial	0	0
Wannon River at Dunkeld	238204	VIC	385	142.34	-37.63	0	High	8	4
Woady Yaloak River Pitfield	234200	VIC	317	143.59	-37.81	0	Moderate	3	0
Collie River at South Branch	612034	WA	672	116.16	-33.39	0	Moderate	6	0

Dunham River at Dunham Gorge	809321	WA	1637	128.3	-16.19	0	High	8	2
Frankland River at Mount Frankland	605012	WA	4467	116.79	-34.91	0	Perennial	0	0
Harvey River at Dingo Road	613002	WA	148	116.04	-33.09	0	Perennial	0	0
Helena River at Ngangaguringurin	616013	WA	316	116	-32.15	0.6	Perennial	0	0
Kalgan River at Stevens Farm	602004	WA	2433	118	-34.89	0	Perennial	0	0
Lefroy Brook River at Rainbow Trail	607013	WA	249	116.02	-34.43	0	Perennial	0	0
Lennard Brook at Molecap Hill	617165	WA	58	115.92	-31.38	25.9	Perennial	0	0
Margaret River at Willmots Farm	610001	WA	435	115.58	-33.56	0	Moderate	3	0
Yale Flat Creek at Woonanup	603190	WA	53	117.29	-34.7	0	Moderate	6	0

1110

1111

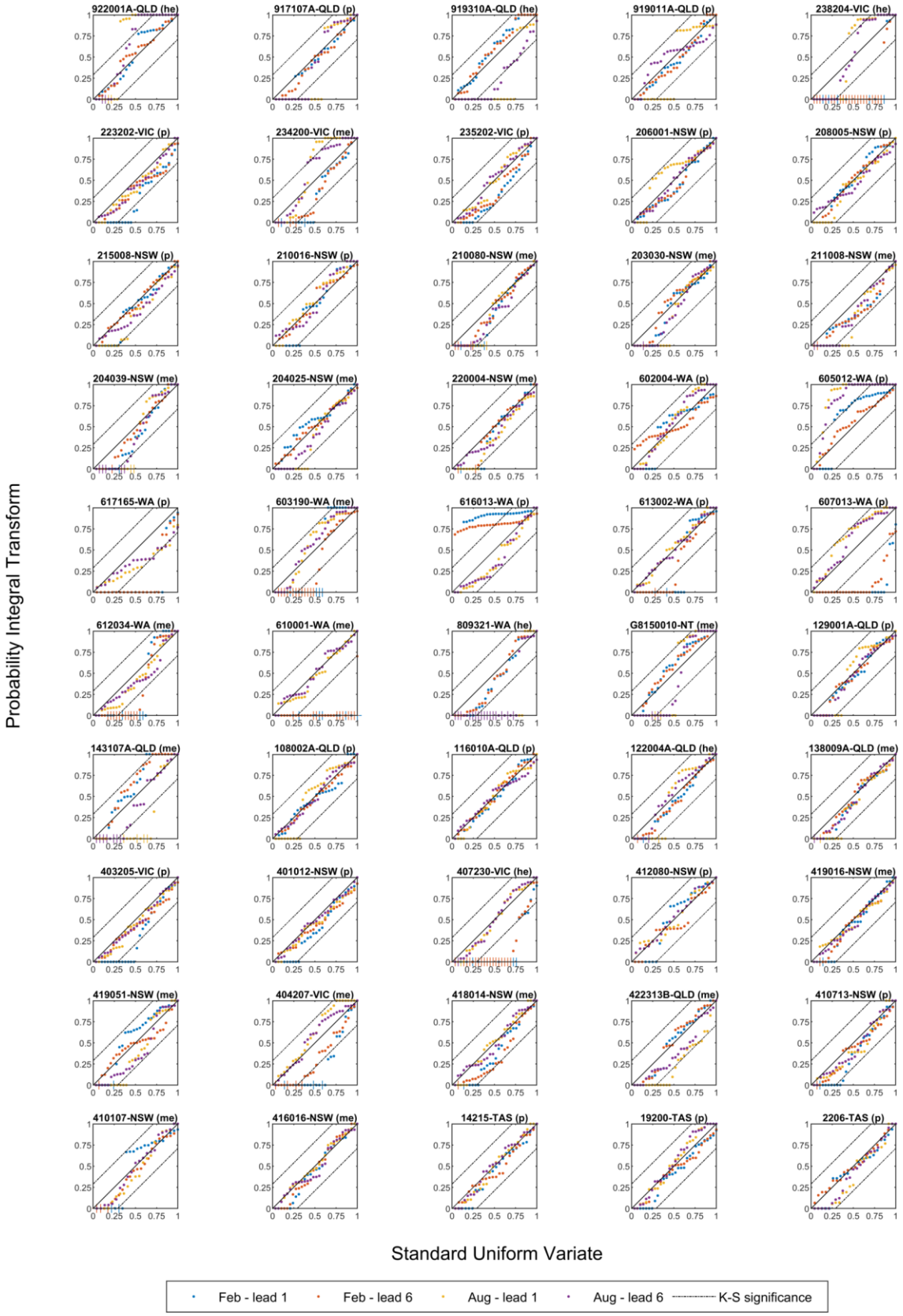


Figure S1 PIT-uniform probability plots for ESP forecasts for selected months and lead times catchments. Degree of ephemerality is denoted by (p)-perennial; (me)-moderately ephemeral; (he)-highly ephemeral.

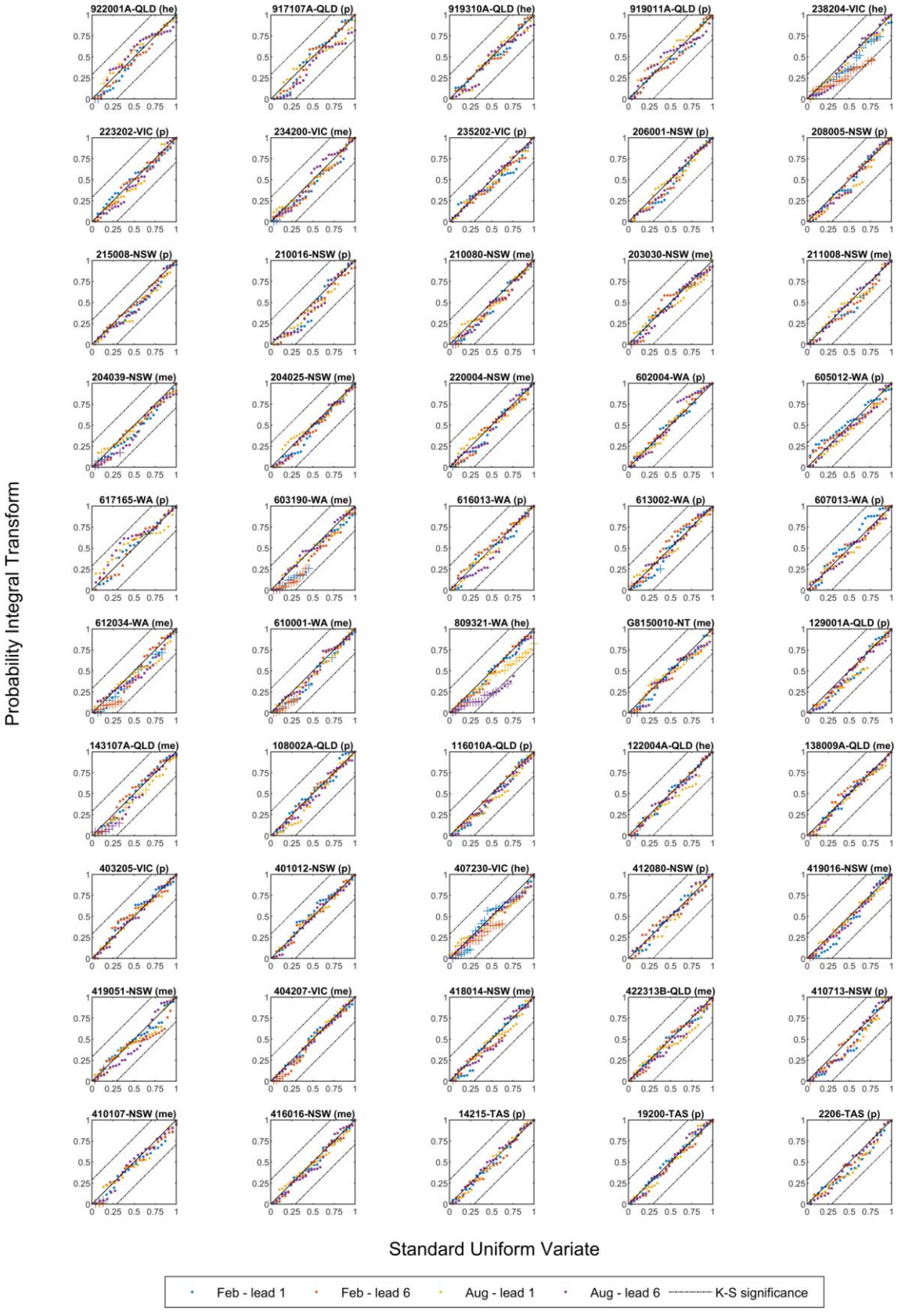


Figure S2 PIT-uniform probability plots for FoGSS forecasts for selected months and lead times catchments. Degree of ephemerality is denoted by (p)-perennial; (me)-moderately ephemeral; (he)-highly ephemeral.

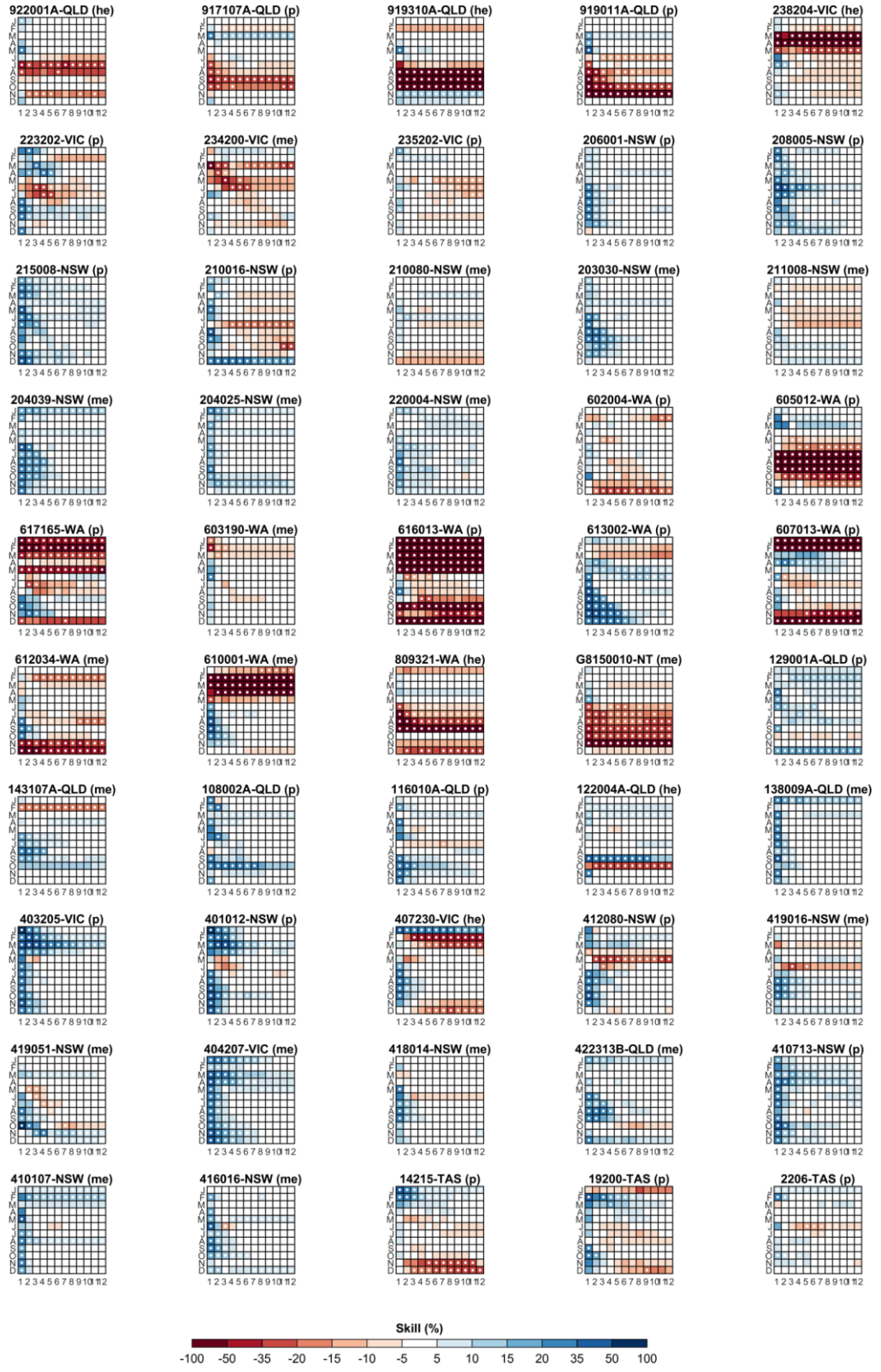
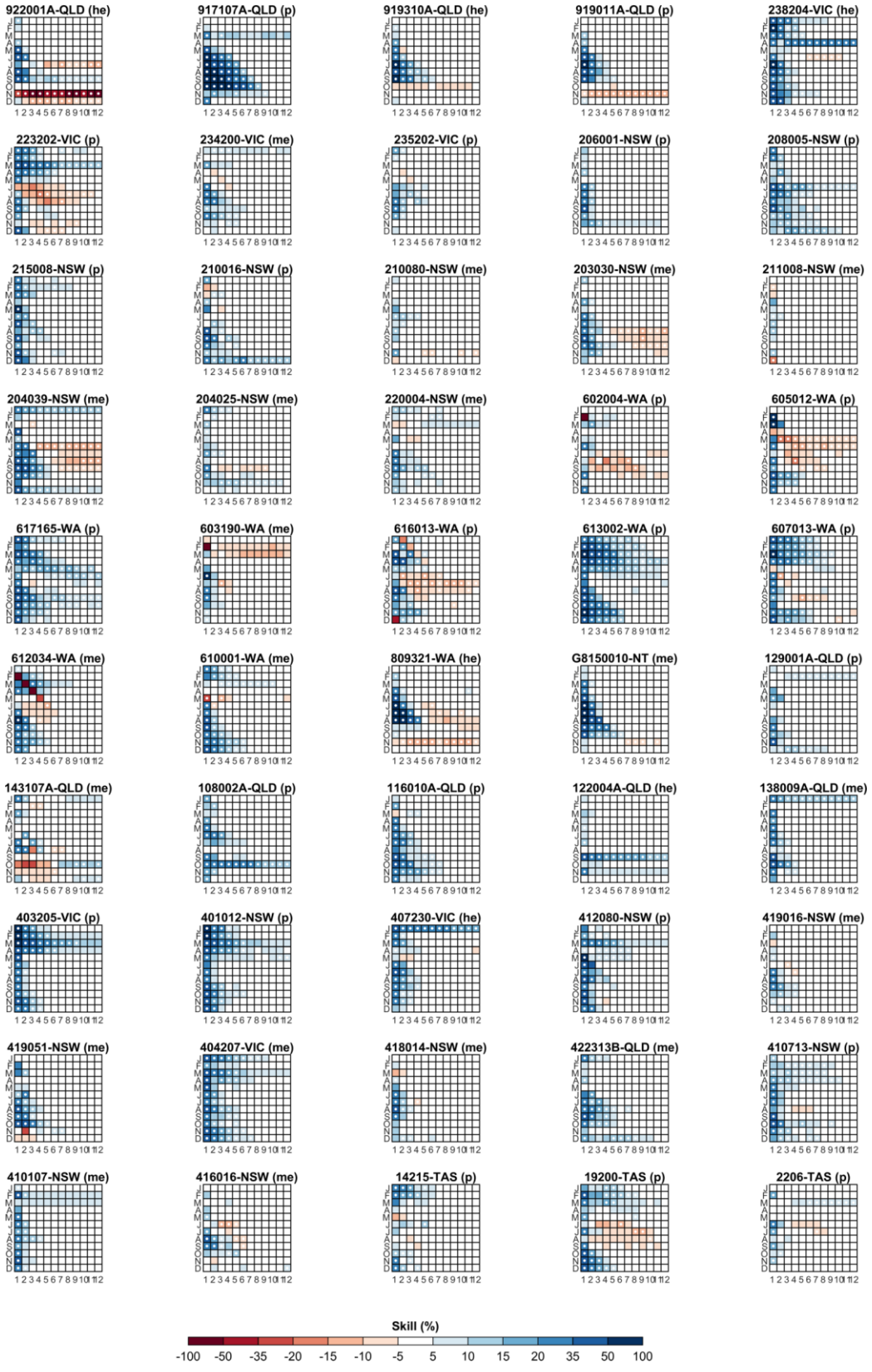


Figure S3 Continuous ranked probability skill scores for ESP forecasts of target months for all catchments. White points indicating statistical significance at the 5% level under bootstrapping (see text for details). Degree of ephemerality is denoted by (p)-perennial; (me)-moderately ephemeral; (he)-highly ephemeral.



1125

1126 Figure S4 Continuous ranked probability skill scores for FoGSS forecasts of target months
1127 for all catchments. Degree of ephemerality is denoted by (p)-perennial; (me)-moderately
1128 ephemeral; (he)-highly ephemeral.