Check for updates

AGU100 ADVANCING EARTH AND SPACE SCIENCE

Water Resources Research

RESEARCH ARTICLE
10.1029/2019WR026128

Key Points:
- Current hydrological error models cannot handle cases where >50% flow is zero
- We introduce a new error modeling method that uses data censoring of simulations and observations
- Our model produces reliable predictions even in highly ephemeral catchments

Supporting Information:
- Supporting Information S1

Correspondence to:
Q. J. Wang,
quan.wang@unimelb.edu.au

Citation:
Wang, Q. J., Bennett, J. C., Robertson, D. E., & Li, M. (2020). A data censoring approach for predictive error modeling of flow in ephemeral rivers. *Water Resources Research*, *56*, e2019WR026128. https://doi.org/10.1029/2019WR026128

Received 8 AUG 2019
Accepted 25 DEC 2019
Accepted article online 3 JAN 2020

A Data Censoring Approach for Predictive Error Modeling of Flow in Ephemeral Rivers

Quan J. Wang[1], James C. Bennett[2,3], David E. Robertson[2], and Ming Li[4]

[1]Department of Infrastructure Engineering, The University of Melbourne, Parkville, Victoria, Australia, [2]CSIRO Land and Water, Clayton, Victoria, Australia, [3]Institute for Marine and Antarctic Studies, University of Tasmania, Hobart, Tasmania, Australia, [4]Floreat, Western Australia, Australia

Abstract Flow simulations of ephemeral rivers are often highly uncertain. Therefore, error models that can reliably quantify predictive uncertainty are particularly important. Existing error models are incapable of producing predictive distributions that contain >50% zeros, making them unsuitable for use in highly ephemeral rivers. We propose a new method to produce reliable predictions in highly ephemeral rivers. The method uses data censoring of observed and simulated flow to estimate model parameters by maximum likelihood. Predictive uncertainty is conditioned on the simulation in such a way that it can generate >50% zeros. Our method allows the setting of a censoring threshold above zero. Many conceptual hydrological models can only approach, but never equal, zero. For these hydrological models, we show that setting a censoring threshold slightly above zero is required to produce reliable predictive distributions in highly ephemeral catchments. Our new method allows reliable predictions to be generated even in highly ephemeral catchments.

Plain Language Summary Many rivers cease to flow at various times. These rivers are difficult to model well, meaning that the models have a high level of uncertainty. There are no existing methods to correctly quantify the uncertainty in models of rivers that cease to flow >50% of the time. We propose a new method that can quantify the uncertainty of these models, even for rivers that cease to flow very often.

1. Introduction

Ephemeral and intermittent rivers drain at least half the world's land surface (Datry et al., 2017; Tooth, 2000), providing water vital for human needs—particularly agriculture—and for ecosystems. They are particularly prevalent in drylands, but they also occur in areas of moderate and high rainfall, notably in headwater catchments (Snelder et al., 2013). Hydrological modeling of these rivers poses a significant challenge, as there is considerable uncertainty in flow predictions (Costigan et al., 2017). Reliable quantification of predictive uncertainty is thus particularly important for ephemeral and intermittent rivers (henceforth simply *ephemeral rivers*) (Smith et al., 2015). While there is considerable literature on quantification of uncertainty in hydrological predictions of perennial rivers (see review by Kavetski, 2019), there are very few published studies for ephemeral rivers.

An important exception is Smith et al. (2010), who highlight the challenges in quantifying predictive uncertainty for ephemeral rivers. They point out that ignoring the presence of zeros undermines two common assumptions of hydrological error models, namely, that residuals will be symmetrically and normally distributed (usually after transformation). To address this problem, they introduced a Bayesian method for parameter inference of hydrological and error models. Within their method, they formulated a likelihood function that is conditioned on whether observed flow was greater than zero or equal to zero. They used this method to demonstrate a fundamental principle of likelihood-based error models in ephemeral rivers: optimal hydrological and error model parameters can only be estimated if zero values are accounted for in the likelihood.

Smith et al.'s (2010) study was aimed at developing a likelihood for ephemeral rivers to optimize hydrological models, rather than for use in predictions. For generating predictions in ephemeral rivers, their method has two limitations: (1) their treatment of zeros in the likelihood estimation is conditioned on observations (Oliveira et al., 2018)—for example, they calculate the "probability of a zero error given a zero observation"—and (2) their error model is incapable of generating >50% zeros, meaning their model cannot produce reliable predictive distributions in highly ephemeral catchments (McInerney et al., 2019), as we demonstrate in section 3.

Conditioning the treatment of zeros on observations is problematic when the model is to be used for prediction: because a future incidence of a zero observation cannot be known, the predictive distribution cannot be correctly conditioned. In previous work, we and others have built on Smith et al.'s (2010) findings with error models that account for zeros in observations, but for which the likelihood is conditioned only on simulations—that is, these models can be used in prediction (Ammann et al., 2019; Bennett, Wang et al., 2016; Li et al., 2013; Li et al., 2017). However, these error models suffer from the second limitation described above: they cannot generate >50% zeros. This is because the original model simulations are treated as the medians of the prediction distributions under the commonly assumed error model. Addressing this limitation is the major aim of this study.

We introduce a maximum likelihood method for parameter estimation of hydrological and error models that can be applied to all catchments, regardless of the degree of ephemerality. To enable the method to be used in prediction, we condition the predictive error distribution on simulated flow from the hydrological model. For intermittent and ephemeral rivers, we assume that both simulated and observed flow may be of zero value, either at the same time or at different times. Thus, the zero threshold needs to be considered in the predictive error distribution both as a condition (simulated flow) and as an outcome (to be observed).

The most direct approach is to use a discrete-continuous predictive error distribution (following Smith et al., 2010), which can be broken down into four cases. Denote observed flow $q(t)$ and simulated flow $\tilde{q}(t)$ at time t. The four cases are the following:

- case = 1 when both $q(t) > 0$ and $\tilde{q}(t)>0$;
- case = 2 when $q(t) = 0$ but $\tilde{q}(t)>0$;
- case = 3 when $q(t) > 0$ but $\tilde{q}(t) = 0$; and
- case = 4 when both $q(t) = 0$ and $\tilde{q}(t) = 0$.

While it is theoretically possible to construct component probability density or cumulative distribution functions for the four cases, it can be quite tedious to do so in practice. This is especially apparent when considering that the distribution functions for the four cases need to satisfy certain relationships. Assuming a monotonically increasing function between the expected value of $q(t)$ and the simulated value $\tilde{q}(t)$, a few of the relationships are as follows:

- For case = 1 and case = 2

$$\int_{0+}^{\infty} \mathrm{pdf}\{q(t)|\tilde{q}(t)\}\mathrm{d}q(t) + \mathrm{prob}\{q(t) = 0|\tilde{q}(t)\} = 1 \tag{1}$$

- For case = 1 and case = 3

$$\int_{Q}^{+\infty} \mathrm{pdf}\{q(t)|\tilde{q}(t)\}\mathrm{d}q(t) > \int_{Q}^{+\infty} \mathrm{pdf}\{q(t)|\tilde{q}(t) = 0\}\mathrm{d}q(t) \tag{2}$$

- For case = 2 and case = 4

$$\mathrm{prob}\{q(t) = 0|\tilde{q}(t)>0\} < \mathrm{prob}\{q(t) = 0|\tilde{q}(t) = 0\} \tag{3}$$

- For case = 3 and case = 4

$$\int_{0+}^{\infty} \mathrm{pdf}\{q(t)|\tilde{q}(t) = 0\}\mathrm{d}q(t) + \mathrm{prob}\{q(t) = 0|\tilde{q}(t) = 0\} = 1 \tag{4}$$

Equations (1) and (4) simply state that the component probabilities need to add to one. Equations (2) and (3) result from the monotonically increasing function.

It is a difficult task to discretely specify the distribution functions for the four cases while satisfying the relationships. The difficulty is exacerbated when some of the cases may have too few data points to be robustly modeled separately.

To resolve these difficulties, we propose a data censoring approach. We treat the zero values of simulated and observed flow as censored data, with unknown exact values equal to or below zero. As we will show in section 2, such an approach makes it possible to use a simple predictive error model to define all the four cases in a mathematically coherent manner. The relationships in equations (1)–(4) are automatically satisfied. Unlike mixed discrete-continuous distributions, data censoring obviates the need for additional distribution(s) to describe the incidence of zero flow in observations or simulations. As a result, data for the four cases can be used jointly, and therefore efficiently, to estimate the parameters of the predictive error model.

We will show that our new method can produce predictive distributions with >50% zeros and is able to handle cases where $\widetilde{q}(t) = 0$. We note, however, that handling cases where $\widetilde{q}(t) = 0$ is of very limited practical value, because many conceptual hydrological models are not able to produce zero flow. To address this issue, we generalize our censoring approach to allow us to set a censoring threshold above zero. We demonstrate that setting a censoring threshold slightly above zero is required to produce consistently reliable predictions with a hydrological model that cannot produce zero values.

The paper is structured as follows. We describe our new method in section 2. We compare the new method to two existing alternatives—one which does not account for zeros at all, and one which accounts only for zeros in observed data—and these experiments are described in section 3. Catchments and data are described in section 4, and verification methods are detailed in section 5. We describe a method to select an appropriate censoring threshold in section 6, before describing the results of our experiments in section 7. We discuss our findings in section 8 and summarize our study in section 9.

2. Predictive Error Model for Ephemeral Catchments

2.1. Hydrological Model

Hydrological modeling 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, Robertson et al., 2016; Coron et al., 2012). Many conceptual models use some form of exponential decay to simulate discharge from stores, with the implication that simulated flow can approach, but never equal, zero. GR4J is exceptional in this regard, as it can produce zero flow with certain combinations of parameters and states. In practice, however, it is quite rare for GR4J to produce zero flow. More usually it produces very long flow recessions that approach, but do not equal, zero.

2.2. Treatment of Threshold Data

We treat the zero values of simulated and observed flow as censored data, with unknown exact values equal to or below zero. Data censoring approaches that deal with threshold data have been successfully used in a number of forecasting applications (Li et al., 2019; Messner et al., 2014; Scheuerer & Hamill, 2015; Wang & Robertson, 2011).

For more general application, we allow the censoring threshold to be zero or above zero, and to have different values for observed and simulated flow, q_C and $\widetilde{q}_C$. As noted above, many hydrological models are formulated in such a way that the simulated flow can never reach exactly zero. Treating the simulated flow below a small positive threshold as having a censored value removes this limitation, thus permitting more realistic modeling of ephemeral rivers.

While our method allows us to set different censoring thresholds for simulated and observed flow, for this study we set $q_C = \widetilde{q}_C$ as a simple test case. Setting $q_C \neq \widetilde{q}_C$ has some potential pitfalls, as discussed in section 8.

The choice of censoring threshold influences the performance of GR4J and the predictive error model, so some care is needed in choosing the censoring threshold. We describe a procedure to choose censoring thresholds in section 6.

2.3. Data Transformation

The predictive error distribution of the simulated flow is often skewed and heteroscedastic, with varied characteristics depending on catchments. One way to simplify the problem is to apply a transformation to normalize the marginal distributions of the observed and simulated flow. After the transformation, a relatively simple predictive error distribution form is often reasonable (e.g., Bates & Campbell, 2001; Thyer et al., 2002; Ye et al., 2014; among many others).

In this study, we use the log-sinh transformation (Wang et al., 2012). The log-sinh transformation was originally designed to improve the representation of residuals in cases where predicted variables are positively skewed, and the spread of errors first increases rapidly, and then more slowly. These properties generally hold for predictions of streamflow (Del Giudice et al., 2013; Wang et al., 2012). We discuss the potential use of other transformations with our method in section 8. The observed flow $q(t)$ is transformed to $z(t)$ by

$$z(t) = tf[q(t)] = b^{-1}\log\{\sinh(a + bq(t))\} \tag{5}$$

where a and b are transformation parameters. The variable z can be back-transformed to q by

$$q(t) = tf^{-1}[z(t)] = b^{-1}(\operatorname{arg\,sinh}[\exp\{bz(t)\}]-a) \tag{6}$$

In the back transformation, any values of $z(t) < tf(0)$ are first forced to $z(t) = tf(0)$ to ensure $q(t) \geq 0$. The same relationships are also applied to the simulated flow $\tilde{q}(t)$ and its transformed value $\tilde{z}(t)$. The censoring thresholds of the transformed flow corresponding to q_C and $\tilde{q}_C$ are denoted as z_C and $\tilde{z}_C$, respectively.

Values of the transformation parameters are estimated from observed flow data. The method of maximum likelihood is used to fit a log-sinh transformed normal distribution to the observed flow data including both censored and noncensored data. The distribution has four parameters: two transformation parameters a and b, and mean m_z and variance s_z^2 of the normal distribution.

The likelihood function is given by

$$L\left(a, b, m_z, s_z^2\right) = \prod_{t:q(t)>q_C} \mathrm{J}\{z(t)\rightarrow q(t)\}\phi\left\{z(t)\middle|m_z, s_z^2\right\} \prod_{t:q(t)\leq q_C} \Phi\left\{z_C\middle|m_z, s_z^2\right\} \tag{7}$$

where data points are denoted by $t = 1,\ldots,T$; $\mathrm{J}\{z(t) \rightarrow q(t)\}$ is the transformation Jacobian given by

$$\mathrm{J}\{z(t)\rightarrow q(t)\} = \coth(a + bq(t)) \tag{8}$$

and $\phi(\cdot| \cdot, \cdot)$ and $\Phi(\cdot| \cdot, \cdot)$ are, respectively, the normal probability density and cumulative distribution functions given mean and variance. Time steps of missing observations should be omitted from the likelihood function. The four parameters are estimated by maximizing the log likelihood by a numerical search method.

To transform the simulated flow, we use the same transformation parameter values as estimated from the observed flow data. In sections 2.4 and 2.5, we will need the marginal distribution of the transformed simulated flow. We assume the distribution is normal, with mean $m_{\tilde{z}}$ and variance $s_{\tilde{z}}^2$. We estimate these two parameters by maximizing the likelihood function

$$L\left(m_{\tilde{z}}, s_{\tilde{z}}^2\right) = \prod_{t:\tilde{z}(t)>\tilde{z}_C} \phi\left\{\tilde{z}(t)|m_{\tilde{z}}, s_{\tilde{z}}^2\right\} \prod_{t:\tilde{z}(t)\leq\tilde{z}_C} \Phi\left\{\tilde{z}_C|m_{\tilde{z}}, s_{\tilde{z}}^2\right\} \tag{9}$$

Note that $m_{\tilde{z}}$ and $s_{\tilde{z}}^2$ must be estimated for each set of hydrological model parameters trialed as part of the parameter estimation procedure described in section 2.5.

2.4. Predictive Error Model

We assume a simple predictive error model

$$z(t)\sim\mathrm{N}\left\{\tilde{z}(t), \sigma^2\right\} \tag{10}$$

that states that the transformed observed flow follows a normal distribution with mean equal to the transformed simulated flow and variance σ^2 being constant.

Under the condition that $\tilde{z}(t)$ is above $\tilde{z}_C$ and its value is known, the probability density function of $z(t)$ is simply

$$\mathrm{pdf}\{z(t)|\tilde{z}(t)\} = \phi\{z(t)|\tilde{z}(t), \sigma^2\} \tag{11}$$

Under this condition, the probability of $z(t) \le z_C$ is

$$\mathrm{cdf}\{z_C|\tilde{z}(t)\} = \Phi\{z_C|\tilde{z}(t), \sigma^2\} \tag{12}$$

Under the condition that $\tilde{z}(t)$ is only known to be equal or below $\tilde{z}_C$, the marginal distribution of $\tilde{z}(t)$ (as described in section 2.3) needs to be invoked to derive the probability density function of $z(t)$ as the following.

$$\mathrm{pdf}\{z(t)|\tilde{z}(t){\le}\tilde{z}_C\} = \frac{\int_{-\infty}^{\tilde{z}_C} \mathrm{pdf}\{z(t)|\tilde{z}(t)\}\phi\left\{\tilde{z}(t)|m_{\tilde{z}}, s_{\tilde{z}}^2\right\}d\tilde{z}(t)}{\int_{-\infty}^{\tilde{z}_C} \phi\left\{\tilde{z}(t)|m_{\tilde{z}}, s_{\tilde{z}}^2\right\}d\tilde{z}(t)} \tag{13}$$

A closed-form solution of this equation is derived and given in section A1.

Under this condition ($\tilde{z}(t){\le}\tilde{z}_C$), the probability of $z(t) \le z_C$ is

$$\mathrm{cdf}\{z_C|\tilde{z}(t){\le}\tilde{z}_C\} = \frac{\int_{-\infty}^{\tilde{z}_C} \mathrm{cdf}\{z_C|\tilde{z}(t)\}\phi\left\{\tilde{z}(t)|m_{\tilde{z}}, s_{\tilde{z}}^2\right\}\mathrm{d}\tilde{z}(t)}{\int_{-\infty}^{\tilde{z}_C} \phi\left\{\tilde{z}(t)|m_{\tilde{z}}, s_{\tilde{z}}^2\right\}\mathrm{d}\tilde{z}(t)} \tag{14}$$

A closed-form solution of this equation is not available. A Monte Carlo integration algorithm for solving this equation is given in section A2.

2.5. Parameter Estimation of Hydrological and Predictive Error Models

Denote the parameter set of a hydrological model being employed as θ. We wish to estimate θ together with the parameter σ^2 in the predictive error model (equation (10)). We use the method of maximum likelihood to do so.

For a given set of values of θ, the hydrological model is run to produce simulated flow values using relevant inputs. After transformation (section 2.3), we have observed and simulated flow values $z(t)$ and $\tilde{z}(t)$, $t = 1,\ldots, T$. The likelihood function is given by

$$L\left(\theta, \sigma^2\right) = \prod_{t:\ \text{case}=1} \mathrm{pdf}\{z(t)|\tilde{z}(t)\} \prod_{t:\ \text{case}=2} \mathrm{cdf}\{z_C|\tilde{z}(t)\} \prod_{t:\ \text{case}=3} \mathrm{pdf}\{z(t)|\tilde{z}(t){\le}\tilde{z}_C\} \prod_{t:\ \text{case}=4} \mathrm{cdf}\{z_C|\tilde{z}(t){\le}\tilde{z}_C\} \tag{15}$$

Time steps of missing observations should be omitted from the likelihood function. The cases in equation (15) are as follows:

- case = 1 when both $z(t)$ and $\tilde{z}(t)$ have known values. The term $\mathrm{pdf}\{z(t)|\tilde{z}(t)\}$ is given by equation (11),
- case = 2 when $z(t) \le z_C$ but $\tilde{z}(t)$ has a known value. The term $\mathrm{cdf}\{z_C|\tilde{z}(t)\}$ is given by equation (12),
- case = 3 when $z(t)$ has a known value, but $\tilde{z}(t){\le}\tilde{z}_C$. The term $\mathrm{pdf}\{z(t)|\tilde{z}(t){\le}\tilde{z}_C\}$is given by equation (13), and
- case = 4 when $z(t) \le z_C$ and $\tilde{z}(t){\le}\tilde{z}_C$. The term $\mathrm{cdf}\{z_C|\tilde{z}(t){\le}\tilde{z}_C\}$ is given by equation (14).

For case = 3 and case = 4, values for $m_{\tilde{z}}$ and $s_{\tilde{z}}^2$ are needed. A method for estimating them is given in section 2.3. For each new set of values of θ, a model simulation run is performed, and new values of $m_{\tilde{z}}$ and $s_{\tilde{z}}^2$ are then estimated. For case = 4, the term $\mathrm{cdf}\{z_C|\tilde{z}(t){\le}\tilde{z}_C\}$ needs to be evaluated only once for each set of values of θ.

A numerical search method (Duan et al., 1993) is used to find the parameters θ and σ^2 that maximize the log-likelihood function of equation (15).

The four cases here correspond to the four cases introduced in section 1. The equivalent relationships of equations (1)–(4) for the more general threshold values used here should also hold. Indeed, it can be shown that the data censoring approach ensures these relationships are satisfied.

We note that our method relies on parameters in the predictive error distribution (equation (10)) and the transformation (equation (5)) to characterize predictive uncertainty. It is not a full Bayesian method that incorporates uncertainty in hydrological and error model parameters. While this is a limitation of our method, in general, the uncertainty due to model parameters is, based on our experience, only a small proportion of total predictive uncertainty of flow. This is particularly true for the long data series that we employ (>20 years; see section 4), and we will show that our method can reliably characterize total predictive uncertainty. Our method is also expected to have the advantage of much greater computational efficiency than Markov Chain Monte Carlo methods required for a full Bayesian analysis.

2.6. Sampling of Predictive Uncertainty

In prediction mode, the predictive error model can be used to generate ensemble members to represent predictive uncertainty.

Given $\tilde{z}(t)$, we generate predictive uncertainty as follows:

1. Condition (1): $\tilde{z}(t)>\tilde{z}_C$. An ensemble member $\tilde{\tilde{z}}(t)$ can be generated by sampling a normal distribution according to equation (10).
2. Condition (2): $\tilde{z}(t)\leq\tilde{z}_C$. A random value of $\tilde{z}(t)$ is sampled from $\mathrm{N}\left(m_{\tilde{z}}, s^2_{\tilde{z}}\right)$ in the range of $\tilde{z}(t)\leq\tilde{z}_C$. Conditioned on this $\tilde{z}(t)$ value, a random value of $\tilde{\tilde{z}}(t)$ is then sampled from the normal distribution of equation (10).

This process is repeated to generate as many ensemble members as desired. For this study, we generate 1,000 ensemble members.

Finally, all ensemble members of $\tilde{\tilde{z}}(t)$ are back-transformed with equation (6) to give the prediction $\tilde{\tilde{q}}(t)$.

3. Error Model Experiments

To test the new method described in section 2, we compare it to two additional approaches commonly used in the literature. We designate the three experiments as follows.

1. n-censored. This is a "no censoring" experiment, using a naive predictive error model that does not account for the presence of zeros or censored values. This is equivalent to considering only case = 1 in our likelihood (section 2.5). Note also that censoring is not used in the estimation of the transformation parameters (equation (7)). For ephemeral rivers, this approach violates the assumptions of normal and symmetrical errors as pointed out by Smith et al. (2010). It has nonetheless been used as a simple "pragmatic" approach in some studies that include ephemeral rivers (McInerney et al., 2017; Woldemeskel et al., 2018; Ye et al., 2014). Predictive uncertainty is generated only with the method used for Condition (1) in section 2.6, including when $\tilde{z}(t)\leq\tilde{z}_C$.
2. o-censored. This experiment treats only observations as censored data, equivalent to using case = 1 and case = 2 in our likelihood (section 2.5). This is analogous to the approach used by Smith et al. (2010) and to error models we have proposed previously (Bennett et al., 2017; Li et al., 2013; Li et al., 2015, 2016). Predictive uncertainty is generated only with the method used for Condition (1) in section 2.6, including when $\tilde{z}(t)\leq\tilde{z}_C$.
3. os-censored. This is the new approach introduced in this study, as described in section 2. It treats both observations and simulations as censored data. Predictive uncertainty is generated with the methods described for both Condition (1) and Condition (2) in section 2.6.

As we note in the introduction (section 1), a fundamental limitation of the n-censored and o-censored experiments is that they cannot generate >50% zeros. We illustrate this with a schematic in Figure 1, for the case where the censoring threshold is the log-sinh transformed value of zero, that is, $z_0 = z_C = \tilde{z}_C = tf[0]$ (equation (5)). In the n-censored and o-censored experiments, only Condition (1) in section 2.6 is used to generate predictive uncertainty (Figure 1a). As the predictive residual distribution is symmetrical, at most ~50% can fall below zero after back-transformation. In the os-censored experiment (Figure 1b), the use of Condition (2) in section 2.6 allows the median prediction $\tilde{z}(t) = \mathrm{M}\left(\tilde{\tilde{z}}(t)\right)$ to fall below z_0, allowing this model to generate >50% of zeros.

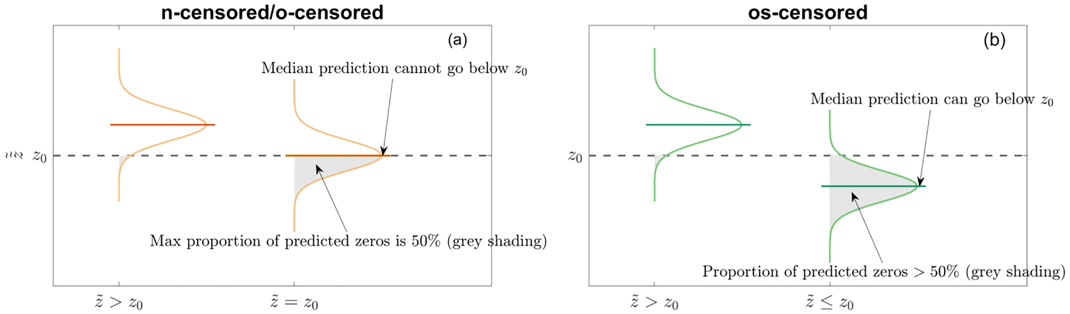
n-censored/o-censored
(a)
Median prediction cannot go below z_0
$\approx z_0$
Max proportion of predicted zeros is 50% (grey shading)
$\tilde{z} > z_0$
$\tilde{z} = z_0$
os-censored
(b)
Median prediction can go below z_0
z_0
Proportion of predicted zeros > 50% (grey shading)
$\tilde{z} > z_0$
$\tilde{z} \leq z_0$

Figure 1. Schematic showing predictive uncertainty under the n-censored/o-censored experiments (a) and the os-censored experiment (b), where $\tilde{z} = \mathrm{M}\left(\tilde{\tilde{z}}\right)$ is the median of the Gaussian predictive distribution and . Each panel shows the treatment of predictive uncertainty when $\tilde{z}>z_0$ and when $\tilde{z}\leq z_0$. Only the os-censored case (Figure 1b) allows >50% zeros to be generated (see text for explanation).

4. Catchments and Data

We choose three catchments to assess the experiments in section 3: Deep River, Fletcher River, and the Ord River. Catchment information is summarized in Table 1, and seasonal distributions of rainfall and flow are summarized in Figure 2. The temperate Deep River is moderately ephemeral, ceasing to flow ~25% of the time, while the tropical Fletcher and Ord Rivers are highly ephemeral, ceasing to flow >50% of the time. Deep River receives winter-dominant (June–November) rainfall and regularly ceases to flow in the late summer and autumn months (January–May). Rainfall in the Fletcher and Ord Rivers is dominated by the monsoon (December–April), and they regularly cease to flow from April to December.

All sites have long (>20 years), high quality flow records. Flow data are taken from the Bureau of Meteorology hydrologic reference stations data set (http://www.bom.gov.au/water/hrs/). This data set provides quality codes, and we set any flow data of quality worse than A ("best available data") to missing. Rainfall and potential evaporation are taken from the gridded AWAP data set (Australian Water Availability Project; http://www.csiro.au/awap/). AWAP produces daily estimates of rainfall interpolated from gauges to a ~5-km grid. AWAP estimates potential evaporation at a monthly time step; we disaggregate these to daily estimates by simple linear interpolation.

5. Model Checking and Verification

As the predictive error model of equation (10) is a critical assumption, we first check if the assumption is reasonable. As parameter estimates of the full models are influenced by estimation methods, including how zero values are treated in the error model experiments (section 3), we check how the error model assumption holds in these experiments. The error model checking should also be useful for explaining predictive performance of the full models.

Ultimately, the purpose of the modeling is for predictions. Therefore, we conduct a thorough evaluation of the performance of the full models in predicting the river flow. Cross-validation is employed in the evaluation to ensure the results are not due to overfitting.

Table 1
Catchment Attributes

Gauge name	Gauge Id	Drainage area (km$^{2)}$	Lat.	Lon.	Zeros (%)	Data period	Missing (%)
Deep River at Teds Pool	606001	471	−34.77	116.62	24.9	01-01-1980 to 31-12-2014	6.8
Fletcher River at Dromedary	803003	66	−17.13	124.99	69.0	01-01-1980 to 13-09-1999	9.8
Ord River at Bedford Downs	809310	546	−17.43	127.60	65.6	01-01-1980 to 31-12-2014	13.9

Note. Dates are formatted as days/month/year.

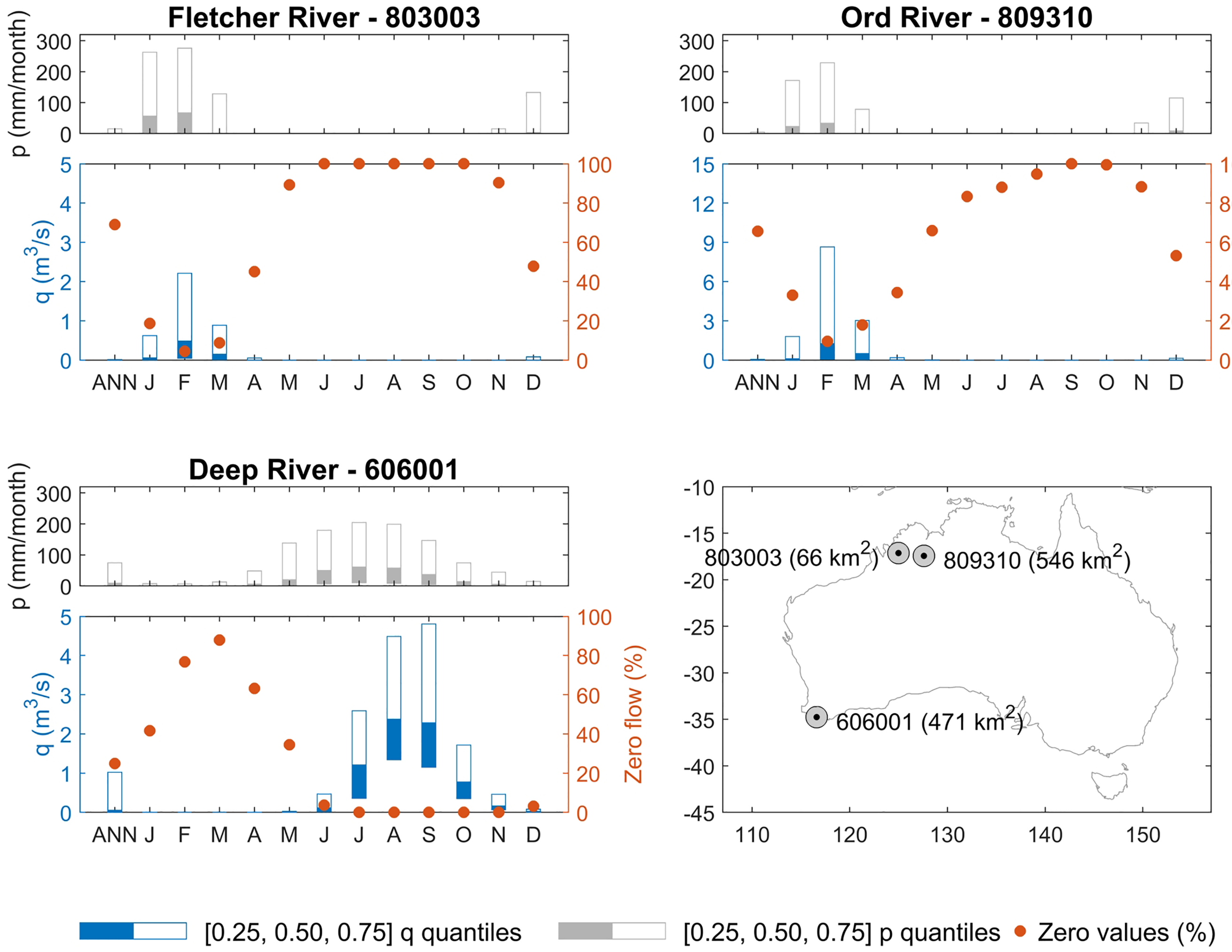

Figure 2. Annual (ANN) and monthly distribution of rainfall (p) and flow (q), including the incidence of zero flow (right-hand axis). Statistics are calculated for the data periods listed for each catchment in Table 1. Catchment locations are given in the bottom right panel.

5.1. Error Model Checking

Because of the presence of threshold values in both $z(t)$ and $\widetilde{z}(t)$, it is not straightforward to check whether the predictive error model of equation (10) is reasonable. Here we check if the predicted residuals are statistically consistent with the observed residuals by allowing for the fact that some of the residuals are affected by the thresholds. Note that this is a check of the theoretical ability of the model to describe residuals in the transform domain, so we do not cross-validate the model in this instance. The following steps are taken:

1. Given the fitted model, we have a set of transformed observations, $z(t)$, and simulations, $\widetilde{z}(t)$. We use these to calculate the residual in the transformed domain:

$$r(t) = \widetilde{z}(t) - z(t). \tag{16}$$

$z(t)$ and $\widetilde{z}(t)$ take different values for each of the censoring experiments, as follows.

a For the n-censored experiment, the values are simply

$$\begin{aligned} z(t) &= z(t) \\ \widetilde{z}(t) &= \widetilde{z}(t) \end{aligned} \tag{17}$$

in all cases, as no censoring is used.

b For the o-censored experiment, $\widetilde{z}(t) = \widetilde{z}(t)$ following equation (17), while $z(t)$ is treated as censored data:

$$\begin{array}{ll} z(t) = z(t) & z(t) > z_c \\ z(t) = \Phi^{-1}\left(\Phi(z_c|\widetilde{z}, \sigma^2) \times U(0,1)\right) & z(t) \leq z_c \end{array} \tag{18}$$

where U(0,1) is a uniform random number between 0 and 1.

c For the os-experiment, $z(t)$ is treated as censored data following equation (18) and $\widetilde{z}(t)$ is also treated as censored data:

$$\begin{array}{ll} \widetilde{z}(t) = \widetilde{z}(t) & \widetilde{z}(t) > \widetilde{z}_c \\ \widetilde{z}(t) = \Phi^{-1}\left(\Phi\left(z_c \middle| m_{\widetilde{z}}, s^2_{\widetilde{z}}\right) \times U(0,1)\right) & \widetilde{z}(t) \leq \widetilde{z}_c \end{array} \tag{19}$$

Note that equation (19) is enacted before equation (18), because equation (18) relies on $\widetilde{z}$.

In other words, $z(t)$ and $\widetilde{z}(t)$ take the form that is seen by the respective likelihood functions used to estimate parameters for each error model experiment.

d Generate histograms of r. This is the frequency distribution of the observed residuals, as seen by the likelihood function for each experiment.

e Overlay the theoretical distribution of residuals estimated by maximum likelihood, given by equation (10).

5.2. Evaluation by Cross Validation

To test the real-world performance of the error models, we evaluate them under a buffered leave-1-year-out cross validation, with a buffer of 4 years. The buffered cross-validation procedure is most easily described using an example. To evaluate predictions for the target year 1990, parameters are estimated using flow data from all years except 1990–1994. These parameters are then used to generate predictions for 1990. The procedure is repeated for each year in the data periods listed for each catchment in Table 1. The 4-year buffer avoids the problem of flow in the target year being informed by flow in the buffer years through the "memory" of GR4J states.

We assess the performance of cross-validated predictions with a range of measures. Model error is calculated for probabilistic predictions using a standardized version of the well-known Continuous Ranked Probability Score (CRPS). For a set of predictions $t = 1,2,...,T$,

$$\mathrm{CRPS} = \frac{\frac{1}{T}\sum_{t=1}^{T} \int_{-\infty}^{\infty} \{F_t(x) - \mathbf{1}[q(t) - x]\}^2 \mathrm{d}x}{\overline{q}} \tag{20}$$

where F_t is the cumulative distribution function (CDF) of the predictive distribution, and **1** is the Heaviside step function. We standardize CRPS by dividing by the mean of observations, $\overline{q}$, to allow comparison of CRPS values between catchments.

We also assess the performance of the underlying deterministic GR4J model. To distinguish between the performance before and after an error model is applied, we term the underlying GR4J model "deterministic simulations" (denoted by $\widetilde{q}(t)$) to distinguish it from the "probabilistic predictions" after the error model is applied (denoted by $\widetilde{\widetilde{q}}(t)$). The deterministic simulation is the median of the predictive distribution, that is, $\widetilde{q}(t) = \mathrm{M}\left(\widetilde{\widetilde{q}}(t)\right)$. We assess the performance of deterministic simulations with the mean absolute error (MAE):

$$\mathrm{MAE} = \frac{\frac{1}{T}\sum_{t=1}^{T} |\widetilde{q}(t) - q(t)|}{\overline{q}} \tag{21}$$

As with CRPS, we standardize MAE by dividing by $\overline{q}$. CRPS and MAE are negatively oriented: smaller values indicate better predictions.

We check predictive reliability with the probability integral transform (PIT). For each prediction, a PIT value is calculated by

$$\pi_t = \begin{cases} F_t(q(t)) & q(t)>0 \\ U[0,1]\times F_t(0) & q(t)=0 \end{cases} \tag{19}$$

If predictions are reliable, the set of PIT values $\{\pi_1, \pi_2, ..., \pi_T\}$ will be uniformly distributed. The treatment of PIT values at $q(t)=0$ is necessary to allow reliable predictions to produce uniformly distributed PIT values when many zero flows occur (Wang & Robertson, 2011). We check the uniformity of PIT values by plotting them as histograms. In addition, we use a less formal measure of reliability: we assess the ability of the predictions to replicate the proportion of zero values (%) in observations.

Finally, we assess sharpness by measuring the average width of prediction intervals of the predictive distributions for the 50% and 90% intervals. As with the error scores, we standardize average width of prediction intervals by dividing by $\overline{q}$.

6. Selection of Censoring Threshold

As we will show in section 7, the selection of a censoring threshold greater than zero is necessary to produce reliable predictions. We initially test six censoring thresholds, $q_C = \widetilde{q}_C = [0, 0.0001, 0.001, 0.01, 0.1, 1.0]$, all in units of m^3/s. As we have already noted, GR4J often cannot produce zeros, a trait it has in common with many hydrological models. Instead, when no rainfall is added to the model over an extended period, simulated hydrographs trail off to infinitesimally small (positive) numbers. The point below which the model is structurally incapable of accurately simulating low flows is readily observable in plots of (empirical) marginal CDFs. We use these to ascertain a suitable positive censoring threshold as follows.

1. Estimate model parameters under o-censoring with $q_C = \widetilde{q}_C = 0$.
2. Generate simulations using parameters estimated in 1.
3. Plot marginal CDFs for observations and simulations generated in 2. To best illustrate the region of interest, we recommend transforming flows with a log transformation and transforming cumulative frequencies with a standard normal variate. Note that simulations and observations should only be taken from the period used in 1 to estimate parameters.
4. Visually assess the marginal CDFs for the point at which low flows in the simulations diverge from low flows in observations. This point (to the nearest order of magnitude) is chosen as the censoring threshold. Where there is ambiguity, we recommend choosing the lowest plausible divergence point: this is to ensure as much data as possible are used to inform inference .

Figure 3 shows these plots for an example calibration period during cross validation. In this example, the divergence points for all three catchments occur at a similar flow: $q_C = \widetilde{q}_C = 0.01$ (all cross-validated calibrations give similar divergence points). For brevity, we will present the remaining results in section 7 for two thresholds: $q_C = \widetilde{q}_C = 0$ and $q_C = \widetilde{q}_C = 0.01$. We provide results for all six censoring thresholds tested in supporting information and discuss sensitivity of our results to the choice of censoring thresholds in section 8.

7. Results

The theoretical ability of each error model to describe transformed residuals is summarized for the highly ephemeral Ord River in Figure 4. The n-censored error model produces transformed residuals that are poorly described by the theoretical Gaussian distribution: the residuals are bimodal and skewed and do not fill the tails of the theoretical distribution. Transformed residuals from the o-censored model are much better represented by the theoretical Gaussian distribution than in the n-censored model but are still imperfect: they are clearly skewed to the right. When $\widetilde{q}_C = 0.0$, the os-censored model behaves similarly to the o-censored model: transformed residuals in both models are skewed to the right. When $\widetilde{q}_C = 0.01$, the os-censored model clearly outperforms the o-censored model: transformed residuals from the os-censored model follow the theoretical Gaussian distribution almost perfectly, while residuals from the o-censored case are still skewed rightward. These findings are reproduced in the highly ephemeral Fletcher River (Figure S2 in the

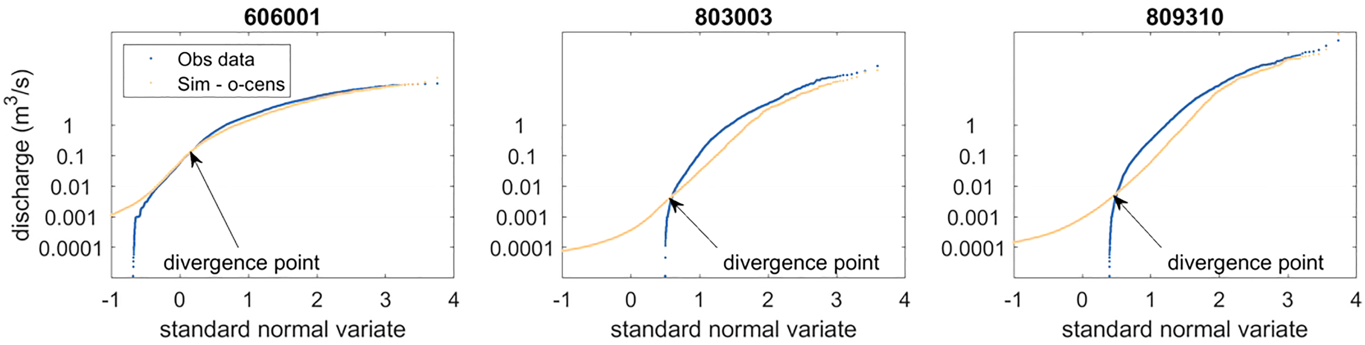

Figure 3. Marginal CDFs of observations and simulations where parameters are estimated with o-censoring. Annotations show the points below which observed and simulated frequency distributions diverge.

supporting information). Conversely, the o-censored and os-censored models perform equally well for the moderately ephemeral Deep River, irrespective of the choice of censoring threshold (Figure S1).

The findings from model checking plots are supported by the PIT histograms of cross-validated predictions, illustrated for the Ord River in Figure 5. Only the os-censored model with $\widetilde{q}_C = 0.01$ produces reliable predictions in highly ephemeral catchments. Both os-censored and o-censored models produce reliable predictions in the moderately ephemeral Deep River, irrespective of censoring threshold. The n-censored predictions are never reliable for any catchment (Figures 5 and S4–S6).

For larger nonzero censoring thresholds ($\widetilde{q}_C >= 0.1$), even predictions for the os-censored threshold become unreliable (see Figures S4–S6). This is due to the large volumes of data subject to censoring, leaving few noncensored data points to inform parameter estimation. For example, for the Ord when $q_C = 1.0$, ~85% of observations are censored. This leads to poor performance at lower flows. In addition, the small set of noncensored data makes the parameter estimation more volatile under cross validation, further contributing to poor performance.

Figure 6 explains why os-censored predictions in highly ephemeral catchments are reliable when $\widetilde{q}_C = 0.01$ but not when $\widetilde{q}_C = 0$. The GR4J model never produces zero flow for the Ord River (Figure 6a). So when $\widetilde{q}_C$

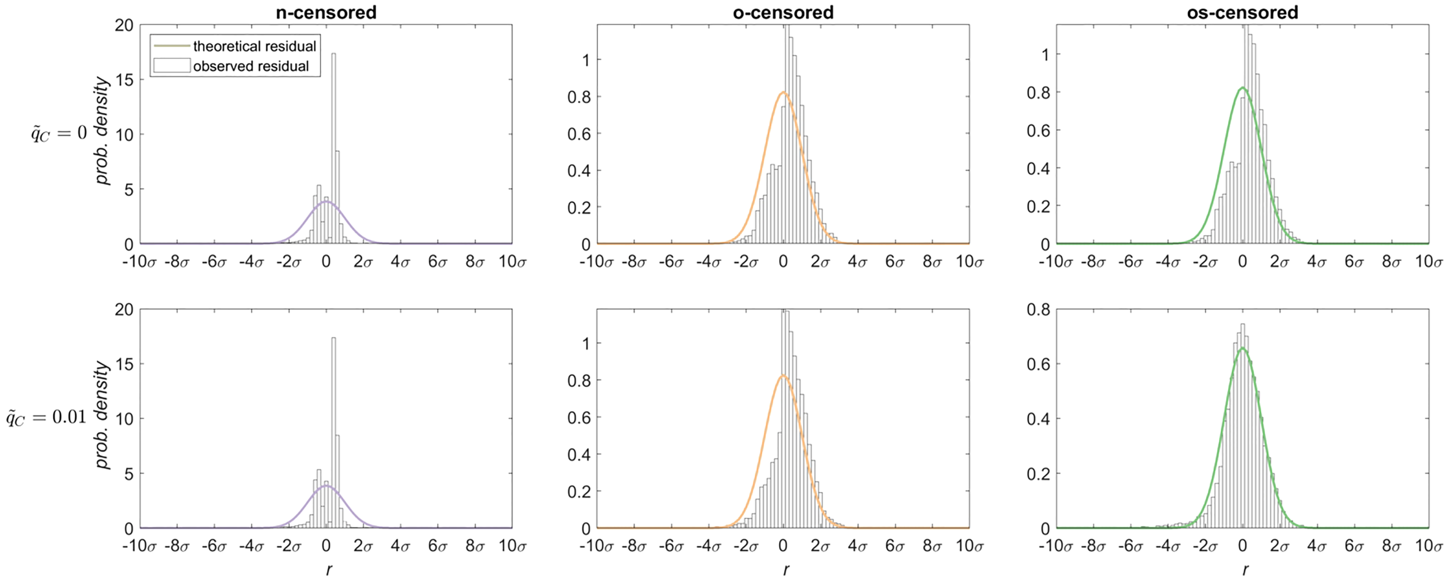

Figure 4. Theoretical and observed residual distributions in the transformed domain for the highly ephemeral Ord River (gauge 809310). Top row shows residuals generated with $\widetilde{q}_C = 0$, and bottom row shows residuals generated with $\widetilde{q}_C = 0.01$. Columns show different censoring experiments. Histograms show frequency distributions of observed residuals in the transformed domain. Lines show theoretical distribution of residuals in the transformed domain. The os-censored case where $\widetilde{q}_C = 0.01$ produces observed residuals that are symmetrically and normally distributed, satisfying the underlying error model assumptions.

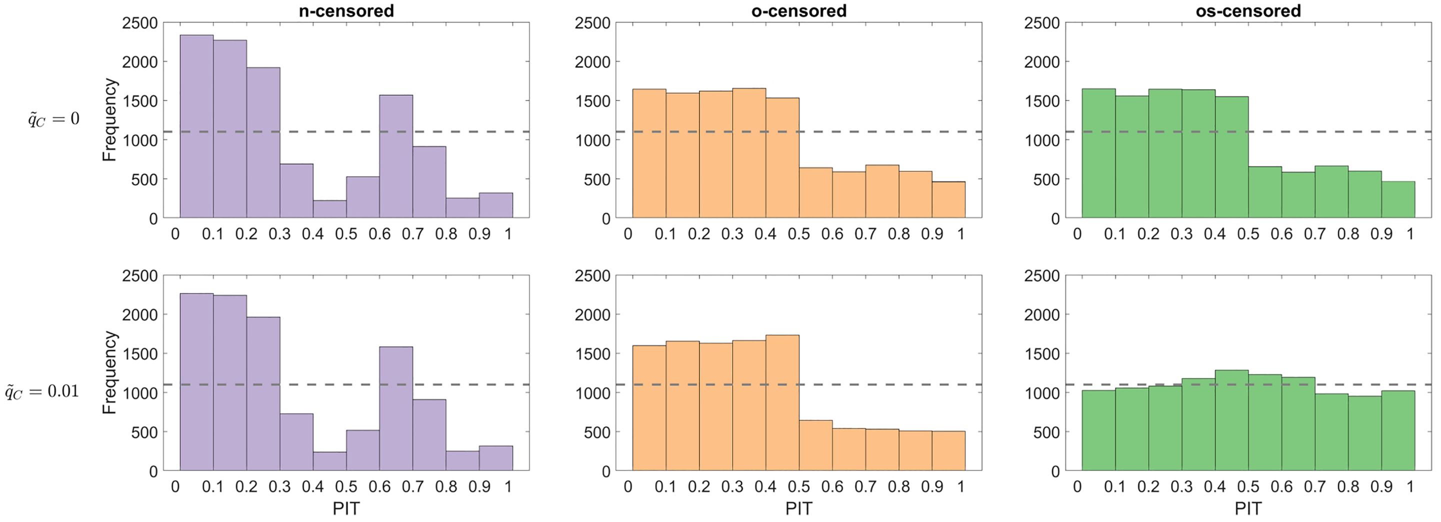

Figure 5. Probability integral transform (PIT) histograms for the highly ephemeral Ord River (gauge 809310). Columns show censoring experiments. Top row shows predictions generated with $\tilde{q}_C = 0$, and bottom row shows simulations generated with $\tilde{q}_C = 0.01$. When PIT values are uniformly distributed (histograms lie along the dashed line), predictions are reliable. Predictions are reliable for the os-censored experiment with $\tilde{q}_C = 0.01$ (bottom right panel).

$= 0$, this means that $\tilde{q}(t) > \tilde{q}_C$ for all t. When $\tilde{q}(t) > \tilde{q}_C$, the o-censored and os-censored models use the same method to generate predictive uncertainty (section 2.6, Figure 1). Thus, the o-censored and os-censored models function very similarly when $\tilde{q}_C = 0$,and neither produces reliable predictions.

Conversely, when $\tilde{q}_C = 0.01$, $\tilde{q}(t) \leq \tilde{q}_C$ for many values of t (Figure 6b). When $\tilde{q}(t) \leq \tilde{q}_C$, the os-censored method treats the simulation as censored data (equation (19)) and generates predictive uncertainty accordingly (section 2.6, Figure 1b). This allows the os-censored model to generate predictive uncertainty distributions with >50% of values below $\tilde{q}_C$, resulting in reliable predictions in highly ephemeral catchments.

Interestingly, a censoring threshold above zero helps to more accurately predict the observed proportion of zero flow (Figure 7). Predictions from the n-censored model strongly underestimate the incidence of zeros in all cases. The o-censored model performs better, but still underestimates the incidence of zeros in all catchments, irrespective of censoring threshold. As with the more formal measures of reliability, the os-censored model is indistinguishable from the o-censored model when $\tilde{q}_C = 0$. However, the os-censored model performs very well when $\tilde{q}_C = 0.01$, closely approaching the observed proportion of zeros in each catchment. This includes the moderately ephemeral Deep River catchment. The improvement when $\tilde{q}_C = 0.01$ can also

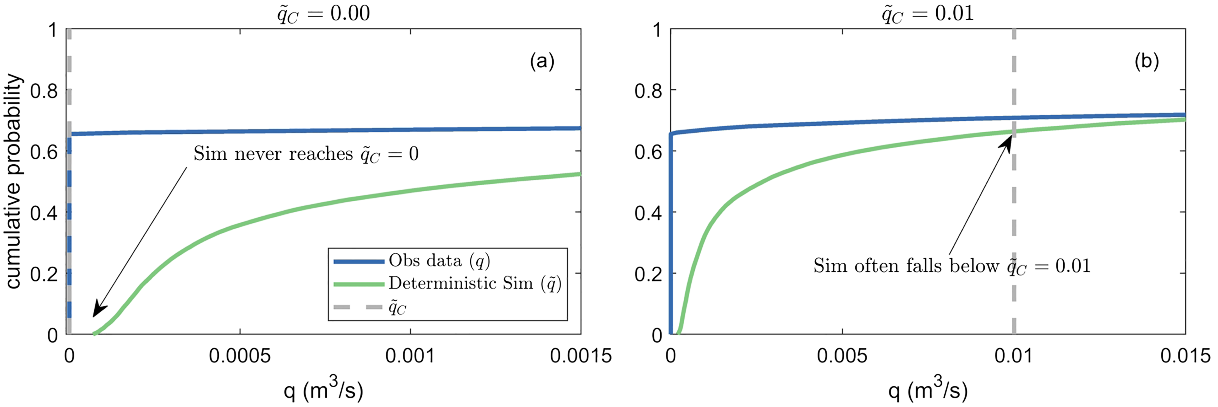

Figure 6. Marginal CDFs of flow for the highly ephemeral Ord River (gauge 809310). Observations (blue line) and deterministic simulations (green line) for the os-censored experiment are shown. Horizontal axes are truncated to focus on very low flow. More than 60% of observations are zero, but when $\tilde{q}_C = 0$ (left panel), simulations never fall to $\tilde{q}_C$. The right-hand panel shows that simulations often fall to/below $\tilde{q}_C$ when $\tilde{q}_C = 0.01$.

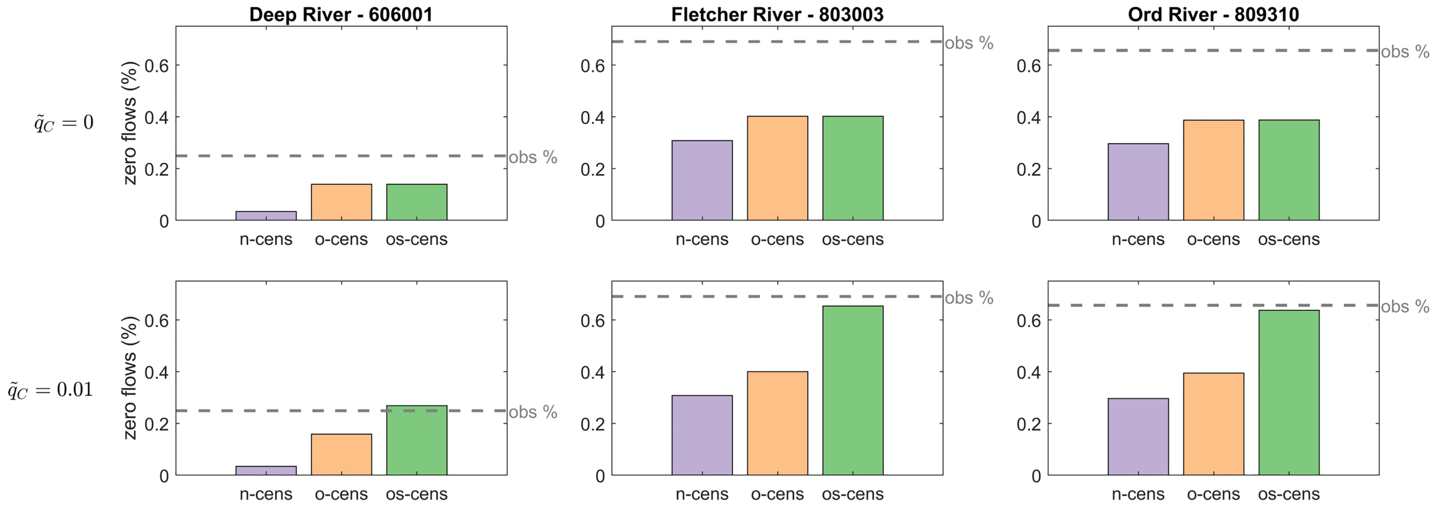

Figure 7. Proportion of zero flow for all gauges (columns). Top row shows predictions generated with $\tilde{q}_c = 0$, bottom row shows predictions generated with $\tilde{q}_c = 0.01$. Colors show different censoring experiments, and dashed line shows the proportion of zeros in the observed record. The os-censored case with $\tilde{q}_c = 0.01$ most closely reproduces the proportion of zeros in all catchments.

be explained by the tendency of GR4J not to produce zero flow, and hence not allowing the os-censored model to take full advantage of the censoring of simulations.

The performance of deterministic simulations from the underlying GR4J model is summarized in Figure 8. In the moderately ephemeral Deep River (gauge 606001), performance of the underlying hydrological model is similar for all error model experiments. In the highly ephemeral Ord (809310) and Fletcher (803003) Rivers, however, the o-censored and os-censored models clearly outperform the n-censored model.

Errors in probabilistic predictions are consistently lower when censored likelihoods are used (o-censored/os-censored) than when no censoring is used (n-censored) (Figure 8, right panels). O-censored predictions have very similar CRPS values to os-censored predictions in all cases, irrespective of censoring threshold.

The accuracy of the underlying GR4J model and the probabilistic predictions are insensitive to the different censoring thresholds we tested, except when $q_C = \tilde{q}_C = 0.0001$ (Figure S7). In all cases where $\tilde{q}_C$>0, the likelihood attempts to match the probability of q_C in observations to the probability of $\tilde{q}_C$ in simulations (Figure S8). (This "probability matching" does not happen for $q_C = \tilde{q}_C = 0$, because in most cases GR4J cannot

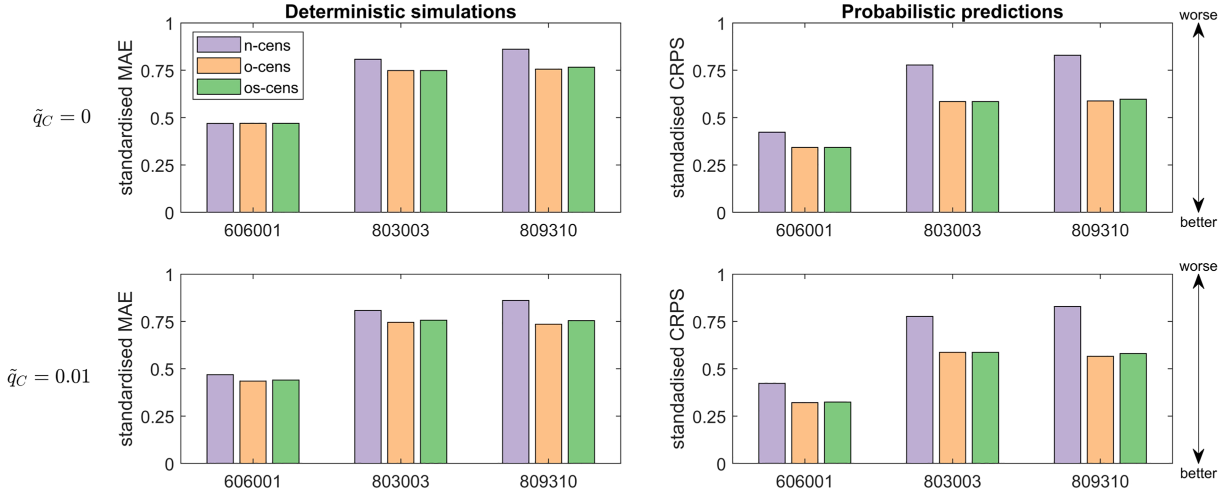

Figure 8. Error scores for all gauges. Left panels show mean absolute error (MAE) of deterministic GR4J simulations, right panels show the continuous ranked probability score (CRPS) of probabilistic simulations. Top row shows predictions generated with $\tilde{q}_c = 0$, bottom row shows predictions generated with $\tilde{q}_c = 0.01$. Colors show different censoring experiments. The n-censored experiment produces larger errors than either the o-censored or os-censored experiments.

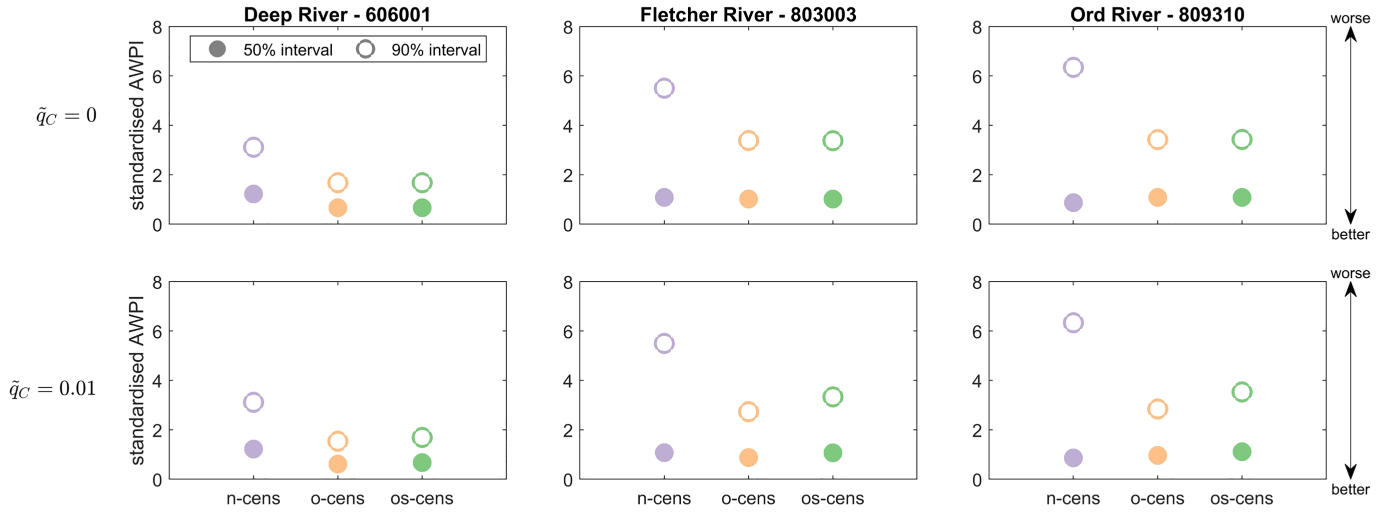

Figure 9. Average width of prediction intervals (AWPI) for 50% (closed circle) and 90% (open circle) intervals, for all gauges (columns). Top row shows predictions generated with $\widetilde{q}_c = 0$, bottom row shows predictions generated with $\widetilde{q}_c = 0.01$. The n-experiment produces wider 90% intervals than either the o-censored or os-censored experiments.

produce zeros.) This becomes problematic at $q_C = \widetilde{q}_C = 0.0001$ for highly ephemeral catchments. The likelihood attempts to force the marginal CDFs shown in Figure 3 to cross at $q_C = \widetilde{q}_C = 0.0001$. This causes the simulated marginal CDF to be forced downward, in turn causing the simulated marginal CDF to poorly match the observed marginal CDF.

As with errors, sharpness of o-censored predictions is very similar to that of os-censored predictions (Figure 9). Both o-censored and os-censored predictions tend to be sharper than n-censored predictions, as we expect given their better representation of predictive uncertainty. The os-censored predictions tend to be slightly less sharp than o-censored predictions when $\widetilde{q}_c = 0.01$. The differences between censored error models and the n-censored model tend to be starkest for the 90% interval. All models show similar sharpness for the 50% interval. Interestingly, the n-censored model tends to produce sharper 50% intervals than the os-censored and o-censored models in the Ord River (gauge 809310). This highlights the importance of considering a range of prediction intervals when assessing sharpness.

The similarity of errors between the o-censored and os-censored predictions illustrates that the major benefit of our method is in the reliability of the predictions. MAE is calculated on the median of the prediction, so does not account for changes in the reliability of the predictive distribution. CRPS does account for reliability but is not strongly sensitive to it. The main reason for this insensitivity is that many of the benefits of the method occur at very low flow. Average measures of error (like MAE and CRPS) tend to emphasize errors at higher flows, as this is when the largest errors occur. Even though residuals at low flows are better represented with os-censoring, the change in the average magnitude of errors is small.

Finally, we illustrate the relative performance of the error model experiments with a time series of the highly ephemeral Ord River (Figure 10). The theoretical failings of the n-censored model manifest in very poor predictive performance, shown by a very poor match between the predicted time series and the observed flow. The o-censored model improves both the underlying hydrological model and the reliability of the predictive distribution, resulting in a much better match between the predicted time series and observations. The further improvements of the os-censored predictions over the o-censored predictions are more difficult to see in Figure 10 but are still evident: slightly wider 90% intervals better encompass observed flow.

8. Discussion

The use of censoring allows us to construct a parsimonious error model that can produce reliable predictions in any catchment, regardless of the degree of ephemerality. This requires the ability to produce >50% zero

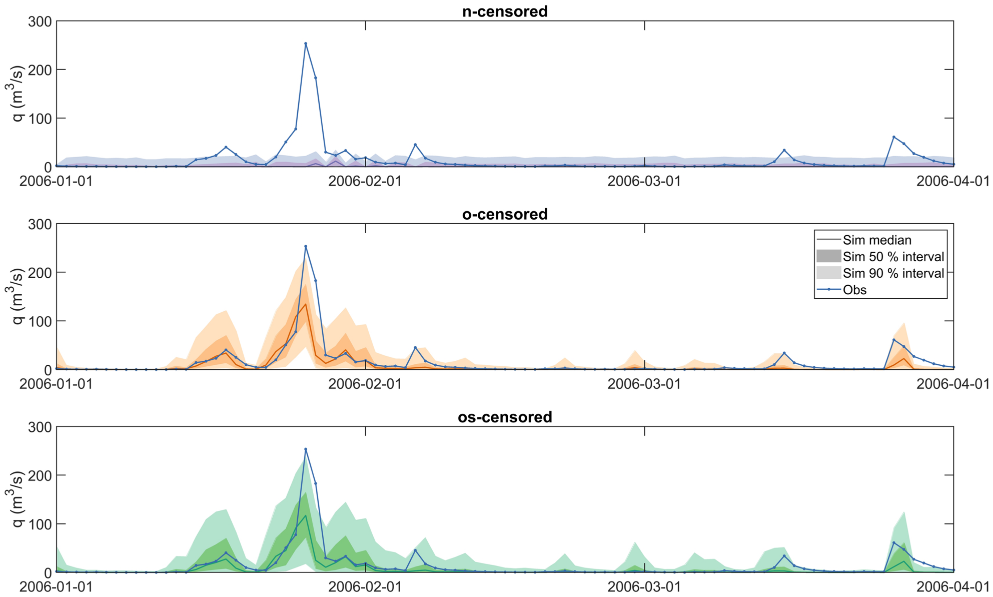

Figure 10. Example simulations for the highly ephemeral Ord River gauge (809310), generated with censoring threshold $\widetilde{q}_c = 0.01$. Top panel shows n-censored model, middle panel shows o-censored model, bottom panel shows os-censored model. The os-censored model produces the most reliable predictions.

flow to enable reliable predictive distributions in highly ephemeral catchments. We show that this is not possible with previously published error models, as embodied by the n-censored and o-censored models. The benefits in reliability that are a feature of our new method do not come at a cost in other performance measures: the os-censored model performs at least as well as the o-censored model in all the measures presented here. In addition, the predictive distribution is conditioned only on simulations, allowing the method to be applied in prediction. Our model could be applied equally well to perennial catchments: in this case censoring will not be invoked, and our model reduces to a simple Gaussian error model after transformation.

For GR4J, the use of a censoring threshold above zero was crucial to ensure reliable predictions in highly ephemeral catchments with the os-censored model. This will be generally true for models that cannot generate zero flow. There are several alternatives to this approach within our method. The first is to use a hydrological model that can generate many zeros (e.g., Ivkovic et al., 2013), allowing the os-censored approach to function when $q_C = \widetilde{q}_C = 0.0$. Such hydrological models are not in common use, and even hydrological models designed to perform well in low flow periods or in ephemeral rivers may not be able to produce zero flow sufficiently frequently or at all (Costelloe et al., 2003; Pushpalatha et al., 2011). Second, accounting for autocorrelation in residuals with an autoregressive (AR) component (e.g., Smith et al., 2015) could push many AR-corrected simulations to zero, possibly allowing the os-censored model to function correctly when $q_C = \widetilde{q}_C = 0.0$. We will explore the use of an AR error model with the likelihood presented in this study in future work.

For our chosen censoring threshold at $q_C = \widetilde{q}_C = 0.01$, most aspects of performance are insensitive within an order of magnitude of this value. That is, we achieved similar performance for $q_C = \widetilde{q}_C = 0.001$ and $q_C = \widetilde{q}_C = 0.1$. However, we find that $q_C = \widetilde{q}_C = 0.0001$ and $q_C = \widetilde{q}_C = 1.0$ can result in poor hydrological model performance and poor reliability, respectively. Accordingly, some care should be taken in choosing censoring threshold by following the method described in section 6. While a value of $q_C = \widetilde{q}_C = 0.01$ was suitable for the three catchments tested here, it is quite possible that appropriate threshold values will vary considerably between catchments and with different hydrological models. In these circumstances, we recommend the threshold be selected for each catchment/model independently.

Our method allows the setting of $q_C \neq \widetilde{q}_C$, for example, $q_C = 0$ and $\widetilde{q}_C = 0.01$. In theory, setting $q_C \neq \widetilde{q}_C$ should produce reliable predictions when $q_C \leq \widetilde{q}_C$ and allows the freedom to adjust censoring thresholds to best suit observations and simulations, respectively. But setting $q_C \neq \widetilde{q}_C$ becomes problematic when $q_C > \widetilde{q}_C$. We fit our transformation to observations (section 2.3), so censoring at q_C means the shape of the transformation is not informed by values below q_C. Thus, when $\widetilde{q}_C$ is set to a value lower than q_C, it is falling in a range of the transformation that may not fit well to observations. We will explore setting $q_C \neq \widetilde{q}_C$ in future work; for now, we recommend setting $q_C = \widetilde{q}_C$.

The ability to choose a censoring threshold above zero may be beneficial for reasons other than achieving reliable predictions. Flow may be difficult to measure accurately at very low values, and stage-discharge rating tables may be subject to considerable uncertainty at very low flow. Rating tables are constructed by fitting curves to gauged stage-discharge data. Some of the mathematical relationships used to describe these curves cannot extrapolate to zero discharge (e.g., linear regressions applied to log-transformed data). For ephemeral rivers, this can mean that the point at which observed flow reaches zero is arbitrarily defined in the rating table: in other words, the probability of zero flow indicated by “observed” flow may not be accurate. In such cases, a censoring threshold above zero, as we have used in this study, could avoid the use of a misleading probability of zero flow. Note that similar reasoning could be applied to uncertain measurements at high flow, and a similar conception of censoring could be applied to an upper bound, although this is outside the scope of the present study.

Our method is somewhat unusual in that it uses a staged parameter estimation procedure: the first stage is to estimate transformation parameters from observations, which are then fixed; the second stage is to estimate the hydrological and error model parameters. Many studies choose instead to fit all parameters jointly (e.g., McInerney et al., 2017; Thyer et al., 2002). The staged approach has the benefit of greater computational efficiency: fewer parameters are estimated at each stage, making the estimation procedure much faster overall. We also prefer the staged approach for conceptual reasons. Hydrological models are structured to replicate heteroscedasticity in flow observations. In this sense, jointly fitting the transformation and hydrological model parameters is likely to result in some interference between transformation and hydrological model parameters. In other words, using the staged approach requires the hydrological model to work harder to match the marginal distribution of observed flow. In principle, however, the likelihood we describe in section 2.5 could be used to jointly estimate all parameters.

Our study confirms that naïve error models that do not account for the presence of zero flow are fundamentally unsuitable for generating predictions in ephemeral rivers. The theoretical failings of these models are well established (Smith et al., 2015; Smith et al., 2010): they violate the assumption of symmetrical, Gaussian residuals when zero flows are present. These theoretical failings have serious practical implications, as we show with the n-censored error model. First, the underlying hydrological models are often not optimal. Second, a naïve error model cannot produce reliable probabilistic predictions, even in moderately ephemeral cases. Following Smith et al. (2010; 2015), we recommend against the use of such naïve error models for ephemeral rivers.

As the log-sinh transformation is applied in our study, there is the question whether the conclusions would be valid had other transformations been applied. Here we draw on the results from the study by McInerney et al. (2019). They compared the use of the log-sinh transformation, the log transformation, and the Box-Cox transformation with a fixed parameter value of 0.2 (BC0.2). When the equivalent of o-censoring was applied, they showed that the different transformations resulted in similar predictive performance in terms of reliability, bias, proportion of zeros, and errors, although BC0.2 led to somewhat sharper distributions. Whether this is also the case for os-censoring needs further investigation.

9. Summary

We present a new method capable of producing reliable predictive distributions even in highly ephemeral catchments with >50% zero flow. Data censoring of both observed and simulated flow is applied when estimating error model parameters by maximum likelihood. A key advantage of our method is the ability to set censoring thresholds above zero. This can compensate for the inability of many conceptual hydrological models to produce zero flows. For highly ephemeral catchments, we show that a censoring threshold set slightly above zero is required to produce reliable predictive distributions with the GR4J hydrological model. We also show that naïve error models that do not account for the presence of zero values in observations of ephemeral rivers are fundamentally unsuitable for producing reliable predictions. We acknowledge that further evaluations of the method in more settings are needed to confirm the generality of our conclusions.

Appendix A

A1. Closed-Form Solution for Equation (13)

Equation (13) reads

$$\mathrm{pdf}\{z(t)|\widetilde{z}(t)\le\widetilde{z}_C\}=\frac{\int_{-\infty}^{\widetilde{z}_C}\mathrm{pdf}\{z(t)|\widetilde{z}(t)\}\phi\left\{\widetilde{z}(t)|m_{\widetilde{z}},s_{\widetilde{z}}^2\right\}\mathrm{d}\widetilde{z}(t)}{\int_{-\infty}^{\widetilde{z}_C}\phi\left\{\widetilde{z}(t)|m_{\widetilde{z}},s_{\widetilde{z}}^2\right\}\mathrm{d}\widetilde{z}(t)} \tag{A1}$$

A closed-form solution for the numerator is derived as follows.

$$\begin{aligned}
&\int_{-\infty}^{\widetilde{z}_C}\mathrm{pdf}\{z(t)|\widetilde{z}(t)\}\phi\left\{\widetilde{z}(t)|m_{\widetilde{z}},s_{\widetilde{z}}^2\right\}\mathrm{d}\widetilde{z}(t)\\
&=\int_{-\infty}^{\widetilde{z}_C}\phi\{z(t)|\widetilde{z}(t),\sigma^2\}\phi\left\{\widetilde{z}(t)|m_{\widetilde{z}},s_{\widetilde{z}}^2\right\}\mathrm{d}\widetilde{z}(t)\\
&=\int_{-\infty}^{\widetilde{z}_C}\frac{1}{\sqrt{2\pi}\sigma}\exp\left[-\frac{\{z(t)-\widetilde{z}(t)\}^2}{2\sigma^2}\right]\frac{1}{\sqrt{2\pi}s_{\widetilde{z}}}\exp\left[-\frac{\left\{\widetilde{z}(t)-m_{\widetilde{z}}\right\}^2}{2s_{\widetilde{z}}^2}\right]\mathrm{d}\widetilde{z}(t)\\
&=\int_{-\infty}^{\widetilde{z}_C}\frac{1}{2\pi\sigma s_{\widetilde{z}}}\exp\left[-\frac{s_{\widetilde{z}}^2+\sigma^2}{2\sigma^2s_{\widetilde{z}}^2}\left\{\widetilde{z}(t)-\frac{s_{\widetilde{z}}^2z(t)+\sigma^2m_{\widetilde{z}}}{s_{\widetilde{z}}^2+\sigma^2}\right\}^2+\frac{\left\{z(t)-m_{\widetilde{z}}\right\}^2}{2\left(s_{\widetilde{z}}^2+\sigma^2\right)}\right]\mathrm{d}\widetilde{z}(t)\\
&=\frac{1}{\sqrt{2\pi\left(s_{\widetilde{z}}^2+\sigma^2\right)}}\exp\left[-\frac{\left\{z(t)-m_{\widetilde{z}}\right\}^2}{2\left(s_{\widetilde{z}}^2+\sigma^2\right)}\right]\frac{\sqrt{\left(s_{\widetilde{z}}^2+\sigma^2\right)}}{\sqrt{2\pi}\sigma s_{\widetilde{z}}}\int_{-\infty}^{\widetilde{z}_C}\exp\left[-\frac{s_{\widetilde{z}}^2+\sigma^2}{2\sigma^2s_{\widetilde{z}}^2}\left\{\widetilde{z}(t)-\frac{s_{\widetilde{z}}^2z(t)+\sigma^2m_{\widetilde{z}}}{s_{\widetilde{z}}^2+\sigma^2}\right\}^2\right]\mathrm{d}\widetilde{z}(t)\\
&=\phi\left\{z(t)\middle|m_{\widetilde{z}},s_{\widetilde{z}}^2+\sigma^2\right\}\int_{-\infty}^{\widetilde{z}_C}\phi\left[\widetilde{z}(t)\middle|\frac{s_{\widetilde{z}}^2z(t)+\sigma^2m_{\widetilde{z}}}{s_{\widetilde{z}}^2+\sigma^2},\frac{\sigma^2s_{\widetilde{z}}^2}{s_{\widetilde{z}}^2+\sigma^2}\right]\mathrm{d}\widetilde{z}(t)\\
&=\phi\left\{z(t)\middle|m_{\widetilde{z}},s_{\widetilde{z}}^2+\sigma^2\right\}\Phi\left\{\widetilde{z}_C\middle|\frac{s_{\widetilde{z}}^2z(t)+\sigma^2m_{\widetilde{z}}}{s_{\widetilde{z}}^2+\sigma^2},\frac{\sigma^2s_{\widetilde{z}}^2}{s_{\widetilde{z}}^2+\sigma^2}\right\}
\end{aligned} \tag{A2}$$

The denominator is simply $\Phi\left(\widetilde{z}_C\middle|m_{\widetilde{z}},s_{\widetilde{z}}^2\right)$, giving the final solution as

$$\mathrm{pdf}\{z(t)|\widetilde{z}(t)\le\widetilde{z}_C\}=\frac{\phi\left\{z(t)\middle|m_{\widetilde{z}},s_{\widetilde{z}}^2+\sigma^2\right\}\Phi\left\{\widetilde{z}_C\middle|\frac{s_{\widetilde{z}}^2z(t)+\sigma^2m_{\widetilde{z}}}{s_{\widetilde{z}}^2+\sigma^2},\frac{s_{\widetilde{z}}^2\sigma^2}{s_{\widetilde{z}}^2+\sigma^2}\right\}}{\Phi\left(\widetilde{z}_C\middle|m_{\widetilde{z}},s_{\widetilde{z}}^2\right)} \tag{A3}$$

A2 Monte Carlo Integration of Equation (14)

Equation (14) reads

$$\mathrm{cdf}\{z_C|\widetilde{z}(t)\le\widetilde{z}_C\}=\frac{\int_{-\infty}^{\widetilde{z}_C}\mathrm{cdf}\{z_C|\widetilde{z}(t)\}\phi\left\{\widetilde{z}(t)|m_{\widetilde{z}},s_{\widetilde{z}}^2\right\}\mathrm{d}\widetilde{z}(t)}{\int_{-\infty}^{\widetilde{z}_C}\phi\left\{\widetilde{z}(t)|m_{\widetilde{z}},s_{\widetilde{z}}^2\right\}\mathrm{d}\widetilde{z}(t)} \tag{A4}$$

Following equation (12), the above equation becomes

$$\mathrm{cdf}\left\{z_C|\widetilde{z}(t)\leq\widetilde{z}_C\right\}=\frac{\int_{-\infty}^{\widetilde{z}_C}\Phi\left\{z_C|\widetilde{z}(t),\sigma^2\right\}\phi\left\{\widetilde{z}(t)|m_{\widetilde{z}},s_{\widetilde{z}}^2\right\}\mathrm{d}\widetilde{z}(t)}{\int_{-\infty}^{\widetilde{z}_C}\phi\left\{\widetilde{z}(t)|m_{\widetilde{z}},s_{\widetilde{z}}^2\right\}\mathrm{d}\widetilde{z}(t)} \tag{A5}$$

For given values of $m_{\widetilde{z}}$ and $s_{\widetilde{z}}^2$, randomly sample a large number (say 1,000) of $\widetilde{z}(t)$ from $\mathrm{N}\left(m_{\widetilde{z}},s_{\widetilde{z}}^2\right)$ in the range of $\widetilde{z}(t)\leq\widetilde{z}_C$. Calculate $\Phi\left\{z_C|\widetilde{z}(t),\sigma^2\right\}$ for each of the sampled $\widetilde{z}(t)$ values. The average of all the $\Phi\left\{z_C|\widetilde{z}(t),\sigma^2\right\}$ values gives an approximate evaluation of $\mathrm{cdf}\left\{z_C|\widetilde{z}(t)\leq\widetilde{z}_C\right\}$.

Acknowledgments

This research has been supported by the Water Information Research and Development Alliance (WIRADA) between the Bureau of Meteorology and CSIRO Land and Water. Streamflow data have been retrieved from the Bureau of Meteorology's hydrological reference station data set, freely available at http://www.bom.gov.au/water/hrs/. Catchment delineations were created with the Bureau of Meteorology's geofabric, freely available from http://www.bom.gov.au/water/geofabric/. Rainfall and potential evaporation data were retrieved from the gridded Australian Water Availability Project (AWAP) data set: http://www.bom.gov.au/jsp/awap/. Derived time series of rainfall and potential evaporation used in this study are available from the authors on request. A full set of Matlab and C++ code used for this study 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. *Hydrology and Earth System Sciences*, *23*(4), 2147–2172. https://doi.org/10.5194/hess-23-2147-2019

Bates, B. C., & Campbell, E. P. (2001). A Markov Chain Monte Carlo Scheme for parameter estimation and inference in conceptual rainfall-runoff modeling. *Water Resources Research*, *37*(4), 937–947. https://doi.org/10.1029/2000wr900363

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

Bennett, J. C., Wang, Q. J., Li, M., Robertson, D. E., & Schepen, A. (2016). 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. https://doi.org/10.1002/2016wr019193

Bennett, J. C., Wang, Q. J., Robertson, D. E., Schepen, A., Li, M., & Michael, K. (2017). Assessment of an ensemble seasonal streamflow forecasting system for Australia. *Hydrology and Earth System Sciences*, *21*(12), 6007–6030. https://doi.org/10.5194/hess-21-6007-2017

Coron, L., Andréassian, V., Perrin, C., Lerat, J., Vaze, J., Bourqui, M., & Hendrickx, F. (2012). Crash testing hydrological models in contrasted climate conditions: An experiment on 216 Australian catchments. *Water Resources Research*, *48*(5), W05552. https://doi.org/10.1029/2011wr011721

Costelloe, J. F., Grayson, R. B., Argent, R. M., & McMahon, T. A. (2003). Modelling the flow regime of an arid zone floodplain river, Diamantina River, Australia. *Environmental Modelling & Software*, *18*(8), 693–703. https://doi.org/10.1016/S1364-8152(03)00071-9

Costigan, K. H., Kennard, M. J., Leigh, C., Sauquet, E., Datry, T., & Boulton, A. J. (2017). Chapter 2.2—Flow regimes in intermittent rivers and ephemeral streams. In T. Datry, N. Bonada, & A. Boulton (Eds.), *Intermittent Rivers and Ephemeral Streams* (pp. 51–78). London, UK: Academic Press. https://doi.org/10.1016/B978-0-12-803835-2.00003-6

Datry, T., Bonada, N., & Boulton, A. J. (2017). Chapter 1—General introduction. In T. Datry, N. Bonada, & A. Boulton (Eds.), *Intermittent Rivers and Ephemeral Streams* (pp. 1–20). London, UK: Academic Press. https://doi.org/10.1016/B978-0-12-803835-2.00001-2

Del Giudice, D., Honti, M., Scheidegger, A., Albert, C., Reichert, P., & Rieckermann, J. (2013). Improving uncertainty estimation in urban hydrological modeling by statistically describing bias. *Hydrology and Earth System Sciences*, *17*, 4209–4225. https://doi.org/10.5194/hess-17-4209-2013

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), 501–521. https://doi.org/10.1007/BF00939380

Ivkovic, K. M., Croke, B. F. W., & Kelly, R. A. (2013). Overcoming the challenges of using a rainfall–runoff model to estimate the impacts of groundwater extraction on low flows in an ephemeral stream. *Hydrology Research*, *45*(1), 58–72. https://doi.org/10.2166/nh.2013.204

Kavetski, D. (2019). Parameter estimation and predictive uncertainty quantification in hydrological modelling. In Q. Duan, F. Pappenberger, A. Wood, H. L. Cloke, & J. C. Schaake (Eds.), *Handbook of hydrometeorological ensemble forecasting* (pp. 481–522). Berlin Heidelberg, Berlin, Heidelberg: Springer. https://doi.org/10.1007/978-3-642-39925-1_25

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. https://doi.org/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. https://doi.org/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. https://doi.org/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. https://doi.org/10.1016/j.jhydrol.2017.10.057

Li, W., Duan, Q., Ye, A., & Miao, C. (2019). An improved meta-Gaussian distribution model for post-processing of precipitation forecasts by censored maximum likelihood estimation. *Journal of Hydrology*, *574*, 801–810. https://doi.org/10.1016/j.jhydrol.2019.04.073

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), 2018WR024148. https://doi.org/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. https://doi.org/10.1002/2016wr019168

Messner, J. W., Mayr, G. J., Wilks, D. S., & Zeileis, A. (2014). Extending extended logistic regression: Extended versus separate versus ordered versus censored. *Monthly Weather Review*, *142*(8), 3003–3014. https://doi.org/10.1175/MWR-D-13-00355.1

Oliveira, D. Y., Chaffe, P. L. B., & Sá, J. H. M. (2018). Extending the applicability of the generalized likelihood function for zero-inflated data series. *Water Resources Research*, *54*(3), 2494–2506. https://doi.org/10.1002/2017WR021560

Perrin, C., Michel, C., & Andréassian, V. (2003). Improvement of a parsimonious model for streamflow simulation. *Journal of Hydrology*, *279*, 275–289. https://doi.org/10.1016/S0022-1694(03)00225-7

Pushpalatha, R., Perrin, C., Le Moine, N., Mathevet, T., & Andréassian, V. (2011). A downward structural sensitivity analysis of hydrological models to improve low-flow simulation. *Journal of Hydrology*, *411*, 66–76. https://doi.org/10.1016/j.jhydrol.2011.09.034

Scheuerer, M., & Hamill, T. M. (2015). Statistical postprocessing of ensemble precipitation forecasts by fitting censored, shifted gamma distributions. *Monthly Weather Review*, *143*(11), 4578–4596. https://doi.org/10.1175/mwr-d-15-0061.1

Smith, T., Marshall, L., & Sharma, A. (2015). Modeling residual hydrologic errors with Bayesian inference. *Journal of Hydrology*, *528*, 29–37. https://doi.org/10.1016/j.jhydrol.2015.05.051

Smith, T., Sharma, A., Marshall, L., Mehrotra, R., & Sisson, S. (2010). Development of a formal likelihood function for improved Bayesian inference of ephemeral catchments. *Water Resources Research*, *46*(12), W12551. https://doi.org/10.1029/2010wr009514

Snelder, T. H., Datry, T., Lamouroux, N., Larned, S. T., Sauquet, E., Pella, H., & Catalogne, C. (2013). Regionalization of patterns of flow intermittence from gauging station records. *Hydrology and Earth System Sciences*, *17*(7), 2685–2699. https://doi.org/10.5194/hess-17-2685-2013

Thyer, M., Kuczera, G., & Wang, Q. J. (2002). Quantifying parameter uncertainty in stochastic models using the Box–Cox transformation. *Journal of Hydrology*, *265*(1–4), 246–257. https://doi.org/10.1016/S0022-1694(02)00113-0

Tooth, S. (2000). Process, form and change in dryland rivers: A review of recent research. *Earth-Science Reviews*, *51*(1), 67–107. https://doi.org/10.1016/S0012-8252(00)00014-3

Wang, Q. J., & Robertson, D. E. (2011). Multisite probabilistic forecasting of seasonal flows for streams with zero value occurrences. *Water Resources Research*, *47*, W02546. https://doi.org/10.1029/2010WR009333

Wang, Q. J., Shrestha, D. L., Robertson, D. E., & Pokhrel, P. (2012). A log-sinh transformation for data normalization and variance stabilization. *Water Resources Research*, *48*, W05514. https://doi.org/10.1029/2011WR010973

Woldemeskel, F., McInerney, D., Lerat, J., Thyer, M., Kavetski, D., Shin, D., et al. (2018). Evaluating post-processing approaches for monthly and seasonal streamflow forecasts. *Hydrology and Earth System Sciences*, *22*(12), 6257–6278. https://doi.org/10.5194/hess-22-6257-2018

Ye, A., Duan, Q., Yuan, X., Wood, E. F., & Schaake, J. (2014). Hydrologic post-processing of MOPEX streamflow simulations. *Journal of Hydrology*, *508*, 147–156. https://doi.org/10.1016/j.jhydrol.2013.10.055