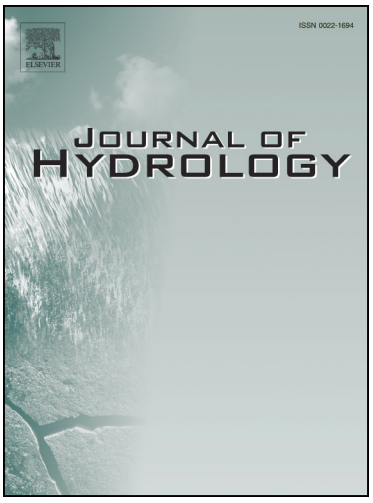
The final publication is available at Elsevier via https://doi.org/10.1016/j.jhydrol.2018.11.047

Accepted Manuscript



Research papers

Efficient treatment of climate data uncertainty in ensemble Kalman filter (EnKF) based on an existing historical climate ensemble dataset

Hongli Liu, Antoine Thiboult, Bryan Tolson, François Anctil, Juliane Mai

PII: S0022-1694(18)30902-8
DOI: https://doi.org/10.1016/j.jhydrol.2018.11.047
Reference: HYDROL 23288

To appear in: *Journal of Hydrology*

Received Date: 29 April 2018
Revised Date: 18 November 2018
Accepted Date: 19 November 2018

Please cite this article as: Liu, H., Thiboult, A., Tolson, B., Anctil, F., Mai, J., Efficient treatment of climate data uncertainty in ensemble Kalman filter (EnKF) based on an existing historical climate ensemble dataset, *Journal of Hydrology* (2018), doi: https://doi.org/10.1016/j.jhydrol.2018.11.047

Efficient treatment of climate data uncertainty in ensemble Kalman filter (EnKF) based on an existing historical climate ensemble dataset

Hongli Liu[1], Antoine Thiboult[2], Bryan Tolson[1], François Anctil[2], Juliane Mai[1]

1. Department of Civil and Environmental Engineering, University of Waterloo, Waterloo, Ontario, Canada.

2. Department of Civil and Water Engineering, Université Laval, 1065 avenue de la Médecine, Québec, Canada

Corresponding author: Hongli Liu.

Tel.: +1 519 888 4567 x37876.

Postal address: Department of Civil and Environmental Engineering, University of Waterloo, 200 University Avenue West, Waterloo, ON, Canada. N2L 3G1

E-mail addresses:

hongli.liu@uwaterloo.ca (H. Liu)

antoine.thiboult.1@ulaval.ca (A. Thiboult)

btolson@uwaterloo.ca (B. A. Tolson)

Francois.Anctil@gci.ulaval.ca (F. Anctil)

juliane.mai@uwaterloo.ca (J. Mai).

Abstract

Successful data assimilation depends on the accurate estimation of forcing data uncertainty. Forcing data uncertainty is typically estimated based on statistical error models. In practice, the hyper-parameters of statistical error models are often estimated by a trial-and-error tuning process, requiring significant analyst and computational time. To improve the efficiency of forcing data uncertainty estimation, this study proposes the direct use of existing ensemble climate products to represent climate data uncertainty in the ensemble Kalman filter (EnKF) of flow forecasting. Specifically, the Newman et al. (2015) dataset (N15 for short), covering the contiguous United States, northern Mexico, and southern Canada, is used here to generate the precipitation and temperature ensemble in the EnKF application. This study for the first time compares the N15 generated climate ensemble with the carefully tuned hyper-parameters generated climate ensemble in a real flow forecasting framework. The forecast performance comparison of 20 Québec catchments shows that the N15 generated climate ensemble yields improved or similar deterministic and probabilistic flow forecasts relative to the carefully tuned hyper-parameters generated climate ensemble. Improvements are most evident for short lead times (i.e., 1-3 days) when the influence of data assimilation dominates. However, the analysis and computational time required to use N15 is much less compared to the typical trial-and-error hyper-parameter tuning process.

KEY WORDS: Climate data uncertainty; hyper-parameter tuning; ensemble Kalman filter (EnKF); short-term ensemble flow forecasting; Newman et al. (2015) dataset

1. Introduction

Ensemble Kalman filter (EnKF) is a sequential data assimilation technique that was proposed by Evensen (1994) as an alternative to the extended Kalman filter. The EnKF uses an ensemble of simulations to represent the distribution of the system state and replaces the covariance matrix by the sample covariance . It is hence well suited to highly nonlinear models (catchment hydrological models in our case) as noted in other studies (McMillan et al., 2013; Reichle and Koster, 2003). The EnKF has many variants, such as using a pair of ensemble Kalman filters (Houtekamer and Mitchell, 1998), a hybrid techniques that combine the EnKF with the 3D variational method (Hamill and Snyder, 2000) or with the variance redactor (Heemink et al., 2001), an ensemble square root filter (Tippett et al., 2003), and a bias-aware retrospective EnKF (Pauwels et al., 2006). These methods all need an ensemble of simulations to represent the state ensemble and model error covariance.

The background error of the state ensemble, before updating, includes internal error and external error. Both are worth considering in the state ensemble generation of the data assimilation. The internal error is introduced by the use of imperfect initial conditions, and the external error refers to the model deficiency (Evensen, 1994). In the EnKF, an ensemble of initial conditions can be generated by adding random noise to the best guess initial conditions or by repeating the warm-up procedure with changed forcing data and/or model parameters (Evensen, 1994; Reichle and Koster, 2003). In the simplest form, the EnKF only accounts for the background error associated with initial conditions (Evensen, 2003; Hamill and Snyder, 2000; Tippett et al., 2003). The external error can be incorporated by either treating the model deficiency as a whole and adding random noise to deterministic model outputs, or explicitly accounting for different sources of

model errors, such as model parameter errors, data errors, and structure errors (Del Giudice et al., 2015; Kumar et al., 2016; Liu and Gupta, 2007). This paper focuses on explicitly accounting for measured climate data uncertainty in the EnKF, which is one part of the external error. For the remainder of this paper, climate data uncertainty and climate ensemble will both refer to the measured historical climate. In contrast, ensemble flow forecasting also can involve climate forecast uncertainty and a corresponding forecasted climate ensemble.

In the EnKF applications of flow forecasting, the most common way of explicitly accounting for climate data uncertainty is treating climate variables as random variables and perturbing climate variables with stochastic errors. In most studies, the error is additive or multiplicative and is assumed to be Gaussian with a predefined constant or proportional variance (Khaki et al., 2017; Rasmussen et al., 2015; Weerts and El Serafy, 2006). Some other probability distributions and stochastic processes are also utilized to generate climate errors (Abaza et al., 2014b; Dunne and Entekhabi, 2006; Eicker et al., 2014; Leisenring and Moradkhani, 2011). Although the predefined error models are easy to construct, they may not reflect the best estimates of the true climate. For example, the common multiplicative stochastic error model approach (e.g., Kavetski et al., 2006a) has the inherent deficiency that it is unable to quantify measurement uncertainty when no rainfall is recorded, which can be especially important in poorly gauged areas (Wright et al., 2017). Another problem is that the distribution variances and stochastic model parameters are often subjectively determined based on the order of magnitude or user's experience of uncertainty (Rasmussen et al., 2015; Reichle et al., 2002).

A solution to reducing the subjectivity in determining climate errors is hyper-parameter tuning, also known in the literature as filter calibration, filter tuning, and EnKF optimization (e.g., Khaki et al., 2017; Reichle and Koster, 2003; Thiboult et al., 2016). In the context of the EnKF, hyper-parameters refer to the parameters of the prior error distributions. Hyper-parameter tuning is a process that recursively tries various sets of hyper-parameter values until the optimal filter performance or forecast performance is found. A typical example is in Reichle and Koster (2003) where a lognormal distributed error is used to perturb measured precipitation values. The standard deviation of the error distribution is determined by trying a selection of standard deviation values until the best filter performance is achieved. The filter performance is assessed by the root mean square error (RMSE) of the aggregated difference between the true state and its EnKF estimate over all catchments. In addition, many studies conduct climate relevant hyper-parameter tuning with other processes to improve the characterization of the background error, such as adjusting the hyper-parameters of system response observation errors (e.g., streamflow errors) (Clark et al., 2008; Reichle and Koster, 2003; Wang et al., 2017), and choosing ensemble size and state variables (Thiboult et al., 2016; Wang et al., 2017).

In addition to the manual hyper-parameter tuning, there are some advanced approaches to reduce the subjectivity of climate uncertainty estimation in flow prediction. The main idea behind these advanced approaches is to infer the climate relevant hyper-parameters with hydrological model parameters based on automatic calibration algorithms. For example, in Bayesian inference, the input error is expressed by an error model. The likelihood function is adjusted to incorporate the input error so that the hyper-parameters can be inferred with hydrological model parameters via Bayesian inference (Del Giudice et al., 2016; Kavetski et al., 2006a, 2006b, Renard et al., 2011,

2010a; Sikorska et al., 2012). Another example is to simultaneously conduct model calibration and data assimilation (e.g., EnKF and particle filter). The hyper-parameters and hydrological model parameters are updated either simultaneously with the states within the assimilation (Moradkhani et al., 2005; Salamon and Feyen, 2010, 2009) or out of each assimilation loop (Vrugt et al., 2005).

These advanced approaches are essentially calibration algorithms that explicitly consider climate uncertainty in parameter inference. In contrast, the previously introduced hyper-parameter tuning is separate from model calibration and is implemented with fixed hydrological model parameters. To our knowledge, the manual hyper-parameter tuning is more popular than any advanced approaches in EnKF based flow forecasting. The main reason is that it is still uncommon for people to explicitly consider climate data uncertainty in model calibration, so the advanced approaches have not been widely applied in practice. Moreover, very few of these advanced approaches have been set up and validated in the real flow forecasting with forecast climate and data assimilation. For instance, the Bayesian inferred climate hyper-parameters are rarely used to generate the climate ensemble for the EnKF. In contrast, in the literature, there are numerous studies adopting the hyper-parameter tuning in EnKF based flow forecasting (e.g., see Abaza et al., 2014a; Li et al., 2014; McMillan et al., 2013; Noh et al., 2014; Reichle et al., 2002; Reichle and Koster, 2003; Thiboult et al., 2016). For example, Thiboult et al. (2016) use hyper-parameter tuning to estimate climate uncertainty because they determined hydrological model parameters before data assimilation without considering climate uncertainty. Therefore, the tuned hyper-parameter approach is taken as the baseline approach to compare our new approach with in this study.

Although hyper-parameter tuning largely solves the subjectivity problem of determining climate errors, it has three limitations. The first is the intensive time and computational cost that users have to spend in the iterative application of data assimilation and forecasting to evaluate filter or forecast performance and find the optimal hyper-parameters for each case study (McMillan et al., 2013; Noh et al., 2014; Slater and Clark, 2006). This issue is due to the ad hoc nature of the hyper-parameter tuning operation. The second limitation is that hyper-parameter tuning mixes the climate uncertainty estimation with the data assimilation and flow forecasting processes. Climate data uncertainty is mostly caused by measurement errors, so its uncertainty estimation depends on measurement errors. However, hyper-parameter tuning determines climate uncertainty after running data assimilation and flow forecasting, the resultant hyper-parameter values may vary with the factors, such as the tuning and data assimilation method, and the climate forecast, which are irrelevant to the climate variable measurement. A consequence of this issue is that sometimes the climate errors are overestimated to compensate other model or initial condition errors and to eventually ensure good filtering and forecast performance (Clark et al., 2008; Evensen, 2007; Thiboult and Anctil, 2015). This is essentially getting the right results for wrong reasons. The third issue is the lack of consideration of the spatio-temporal correlation in generating climate errors. It is easy to understand that accounting for the spatio-temporal correlation gives a better description of the true climate. A better climate ensemble gives a better description of background error and thereby a better filter performance (McMillan et al., 2013; Rasmussen et al., 2015; Reichle and Koster, 2003). In practice, most EnKF applications neglect the climate spatio-temporal correlation (e.g., Abaza et al., 2014; Eicker et al., 2014; Rasmussen et al., 2015; Thiboult and Anctil, 2015; Whitaker and Hamill, 2002). Part of the reason is for

simplicity, but the more important reason is that the spatio-temporal correlation characteristics of the uncertain climates are unknown beforehand and hard to quantify (Rasmussen et al., 2015). In the very few studies that account for the correlation(s), the temporal correlation is typically modelled with an autoregressive model of order one, and the spatial correlation is computed by the nested grid approach or the Fourier transform (Clark et al., 2008; Reichle and Koster, 2003; Tangdamrongsub et al., 2015).

Given these limitations of hyper-parameter tuning, a small number of studies have tried to avoid it by generating a climate ensemble (and a system response observation ensemble) before the EnKF phase. Slater and Clark (2006) and Clark et al. (2006) generate precipitation and temperature ensembles prior to the EnKF, based on a geo-statistical method introduced by Clark and Slater (2006). Huang et al. (2017) directly use the 100 members of a historical ensemble climate dataset developed by Newman et al. (2015) to force their hydrological model in the EnKF. The latter dataset is generated by following the geo-statistical method of Clark and Slater (2006) but making several modifications, among which the foremost is incorporating the temporal correlation in the spatially correlated random field generation (Newman et al., 2015). For short, the Newman et al. (2015) dataset is referred to as N15 in the remainder of this paper. Although N15 has been used in some applications, more precisely, seasonal streamflow simulations (not forecasting), it has not yet been established that it yields better streamflow forecasting than the carefully tuned hyper-parameters based climate ensemble. The answer to this question has wide implications for the applications of the EnKF and its variants. If N15 produces the practically same forecasting results as the carefully tuned hyper-parameters do, then the subjective and arduous hyper-parameter tuning practice can be eliminated. The saved time

can instead be used to further enhance forecast performance in other ways, such as improving the model parameters and structure, and taking into account the other model errors in filter calibration.

Therefore, the objectives of this study are to: (1) compare the climate ensemble generated by N15 with the climate ensemble generated by carefully tuned hyper-parameters, and (2) compare their flow forecast results over a large number of catchments in the EnKF based ensemble flow forecasting. This is the first study to compare the N15 generated ensemble with the carefully tuned hyper-parameters generated climate ensemble in a real flow forecasting framework. The tuned hyper-parameters and corresponding flow forecasting results are taken from Thiboult et al. (2016), and relevant details are provided here in Section 2 and Section 3.

The remainder of the paper is organized as follows. Section 2 describes in detail the EnKF method, forecasting experiments, and two comparative approaches of generating climate ensembles. Section 3 presents the comparison results and discussion of the climate ensembles and the flow forecasts over 20 Québec catchments. Conclusions and future work can be found in Section 4.

2. Methods and Data

2.1. Ensemble Kalman Filter

This section provides a brief summary of the EnKF algorithm. More information about the EnKF equations and mathematical background can be found in Evensen (2003) and Houtekamer and Mitchell (2001). The state vector **X** evolves according to:

$$\mathbf{X}_t^- = \mathbf{M}_t\left(\mathbf{X}_{t-1}^+, \mathbf{U}_t, \theta\right) + \boldsymbol{\eta}_t \quad (1)$$

where $\mathbf{X}^-$ and $\mathbf{X}^+$ represent the prior and posterior estimates of the state, respectively. **M** is the non-linear forward operator forced by the previous state, the climate input **U**, and the model parameter θ. $\boldsymbol{\eta}$ is the model error due to uncertainties in model structure, model parameters, initial conditions, and input data.

The state is transformed to the system response observation **Z** by:

$$\mathbf{Z}_t = \mathbf{H}_t\left(\mathbf{X}_t^-\right) + \boldsymbol{\epsilon}_t \quad (2)$$

where **H** is the observation operator that converts the model state to the observation. $\mathbf{H}(\mathbf{X}^-)$ is the prior estimate of the system response. $\boldsymbol{\epsilon}$ is the response observation error.

When a response observation is available, the model state can be updated as a weighted average between the prior state and the difference between the prior estimate and observation of the system response:

$$\mathbf{X}_t^+ = \mathbf{X}_t^- + \mathbf{K}_t(\mathbf{Z}_t - \mathbf{H}_t\left(\mathbf{X}_t^-\right)) \quad (3)$$

where **K** is the Kalman gain. **K** functions as the weight in a state update and is calculated by:

$$\mathbf{K}_t = \mathbf{P}_t\mathbf{H}_t^T(\mathbf{H}_t\mathbf{P}_t\mathbf{H}_t^T + \mathbf{R}_t)^{-1} \quad (4)$$

where **P** is the covariance of the state error, and **R** is the response observation error covariance.

When solving the Kalman gain, Houtekamer and Mitchell (2001) propose calculating $\mathbf{PH}^T$ and $\mathbf{HPH}^T$ directly from the ensemble members, rather than calculating each element of Equation (4):

$$\mathbf{P}_t\mathbf{H}_t^T = \frac{1}{N-1}\sum_{i=1}^{N}(\mathbf{X}_{t,i}^- - \overline{\mathbf{X}_t^-})[\mathbf{H}(\mathbf{X}_{t,i}^-) - \overline{\mathbf{H}(\mathbf{X}_t^-)}]^T \quad (5)$$

$$\mathbf{H}_t\mathbf{P}_t\mathbf{H}_t^T = \frac{1}{N-1}\sum_{i=1}^{N}[\mathbf{H}(\mathbf{X}_{t,i}^-) - \overline{\mathbf{H}(\mathbf{X}_t^-)}][\mathbf{H}(\mathbf{X}_{t,i}^-) - \overline{\mathbf{H}(\mathbf{X}_t^-)}]^T \quad (6)$$

where N is the ensemble size ($i = 1,\ldots, N$), and

$$\overline{\mathbf{X}_t^-} = \frac{1}{N}\sum_{i=1}^{N}\mathbf{X}_{t,i}^- \quad (7)$$

$$\overline{\mathbf{H}(\mathbf{X}_t^-)} = \frac{1}{N}\sum_{i=1}^{N}\mathbf{H}(\mathbf{X}_{t,i}^-) \quad (8)$$

Since the focus of this research is investigating the influence of different climate ensembles on flow forecasting, it is important to clarify how the change of climate inputs affects the EnKF. Changes in the input **U** will be firstly passed to the prior state $\mathbf{X}^-$ according to Equation (1). Change of state estimate $\mathbf{X}^-$ will then affect the prior estimate of system response $\mathbf{H}(\mathbf{X}_t^-)$ based on Equation (2) and finally affect the state update based on Equation (3).

2.2. Hydrological Model

Thiboult et al. (2016) use 20 hydrological models to account for model structure uncertainty in ensemble flow forecasting, arguing that different hydrological models may compensate each other in terms of the overall forecast performance. Since the focus of this study is on the climate

ensemble, the multi-model approach is not retained to ensure that the different sources of uncertainty are disaggregated. A single hydrological model, GR4J (Génie Rural à 4 paramètres Journalier) model, corresponding to the 5th model of Thiboult et al. (2016), is chosen to provide a fair comparison of the different climate ensembles.

The GR4J model is a daily time-step, lumped four-parameter rainfall-runoff model. It is proposed by Perrin et al. (2003) and has been applied in numerous studies (e.g., Demirel et al., 2013; McInerney et al., 2017; Renard et al., 2010). In GR4J, basin processes are described by a production store and a routing store. The model includes a conceptual representation of the main hydrological processes such as percolation, routing, and groundwater exchange. GR4J takes the precipitation depth and the potential evapotranspiration as input.

In addition, two methods are employed to provide necessary driving forces for GR4J in Thiboult et al. (2016). The precipitation depth is calculated by the two-parameter snow accounting routine Cemaneige (Valéry et al., 2014) that is driven by daily precipitation and air temperature and generates the amounts of rain and snowmelt of the catchment. The potential evapotranspiration is estimated by a conceptual formula proposed by Oudin et al. (2005) based on air temperature and calculated radiation.

There is a total of six parameters in the hydrological model - four in GR4J and two in Cemaneige. The parameter values for each catchment are taken from Thiboult et al. (2016) who determined the optimal parameter sets by minimizing the root mean square error between the simulated and

observed flows over the 1990-2000 period with the shuffled complex evolution calibration method (Duan et al., 1992).

2.3. Research Area and Data

Our research is conducted on the same 20 catchments as Thiboult et al. (2016) that are located in southern Québec, Canada. The 20 catchments have different physiographic characteristics and hydrological responses. Some main characteristics are listed in Table 1.

Table 1. Main characteristics of the 20 Québec catchments. Q and $P$are the observed streamflow and precipitation, respectively (from Table 1 of Thiboult et al. (2016)).

No.	River name	Area (km^2)	River length (km)	Average slope (%)	Mean ann. Q (m^3/s)	Coeff. of variation of Q[1]	Mean ann. P (mm)	Mean ann. Snow (cm)[2]
1	Trois Pistoles	923	52	0.52	18	1.81	1109	382
2	Du Loup	512	45	0.78	10	1.47	1050	378
3	Gatineau	6796	190	0.12	127	1.08	1023	332
4	Dumoine	3743	145	0.13	50	0.81	968	297
5	Kinojevis	2572	83	0.12	39	1.12	921	324
6	Matawin	1383	68	0.29	24	1.11	1025	328
7	Croche	1551	102	0.33	29	1.24	996	360
8	Vermillon	2650	145	0.20	39	1.10	957	312
9	Batiscan	4483	167	0.45	96	1.03	1162	381
10	Sainte Anne	1539	84	0.81	51	1.20	1412	502
11	Bras du Nord	643	77	0.82	19	1.21	1385	499
12	Du loup	767	57	0.78	12	1.27	1020	332
13	Aux Ecorces	1107	54	1.04	28	1.09	1236	450
14	Metabetchouane	2202	155	0.43	48	1.19	1168	420
15	Peribonka	1010	101	0.50	19	1.16	1000	376
16	Ashuapmushuan	15342	342	0.16	300	0.92	984	379
17	Ashuapmushuan	11200	232	0.12	227	0.88	1001	394
18	Au Saumon	586	69	0.65	8	1.36	877	334
19	Mistassini	9534	278	0.20	200	1.08	1004	409
20	Valin	761	59	1.06	24	1.13	1123	453

Note: 1. Coefficient of variation is the ratio of the standard deviation to the mean. It shows the extent of variability relative to the mean for daily flow. 2. Mean annual snow is the average yearly snowfall depth.

Measurements of streamflow, precipitation and maximum and minimum temperatures are provided by the Direction de l'Expertise Hydrique. Precipitation and temperature data are gridded and generated by Kriging interpolation over a 0.1° grid cell given station measurements. According to the Québec climate monitoring program (Bergeron, 2016), precipitation is measured by tipping bucket gauge, heated tipping bucket gauge, or weighting rain gauge. The daily precipitation and temperature measurements are interpolated based on 392 meteorological stations over the domain from $43^{\circ}N$ to $53^{\circ}N$ and from $55^{\circ}W$ to $81.5^{\circ}W$. The 20 catchments of Table 1 are situated in the region where the station density is relatively high. The Kriging error is less than $0.1\,mm/d$ for precipitation and is around $0.1℃$ for both the maximum and the minimum temperatures.

Forecast precipitation and maximum and minimum temperatures are retrieved from the THORPEX interactive grand global ensemble (TIGGE) database. The raw forecast data are at 0.5° and 6-hour resolution. They are downscaled from 0.5° to 0.1° by a bilinear interpolation and aggregated from 6-hour to daily time step to improve spatial resolution and meet the time step requirement of the hydrological model. The forecast climate has 50 members, and the forecast lead time is 9 days. In model application, the forecast climate is lumped to the catchment scale by calculating the average of all the grid cells within the catchment.

2.4. Forecasting Experiment

In the forecasting phase, simulation, data assimilation, and forecasting alternate as follows: (1) the model is forced with the measured climate up to the first day *t* of the forecast, (2) the state

estimates are updated based on the measured flow with the EnKF, and (3) the model is forced with meteorological forecasts to generate hydrological ensemble flow forecasts until *t*+9 days.

In Thiboult et al. (2016), the simulation and forecasting periods are November 1, 2003 to October 31, 2008 and November 1, 2008 to December 1, 2010, respectively. In the simulation period, the model is started with the same initial states (e.g., water levels of two stores) on November 1, 2003 and evolves until October 31, 2008 with the measured climate. In simulation, the EnKF is implemented to generate multiple state conditions as a consideration of the initial condition error for forecasting. In data assimilation, Thiboult et al. (2016) explicitly addresses the model errors from initial conditions and climate input data (i.e., precipitation and temperature) as well as the flow observation error in the EnKF, while it does not account for the model structure and parameter errors in an explicit manner. The EnKF ensemble size is 50. State variables are daily updated.

A forecasting system can be identified by the climate forecast and a collection of settings for the hydrological model and the data assimilation technique. Since the EnKF is the focus of this study, the settings for the climate forecast and hydrological model are identical from one forecasting system to another, and only the EnKF implementation is varied to compare the performances of three forecasting systems. The three systems are differentiated by their climate ensembles used in the EnKF. Two climate ensembles are generated by the carefully tuned hyper-parameters, while the third climate ensemble is generated from the N15 dataset. The detailed climate ensemble generation processes and their corresponding forecasting systems are explained below.

2.4.1. Traditional Ensemble Generation

This study uses two forecasting systems of Thiboult et al. (2016): system H and system H'. More precisely, the subsets of systems H and H' since a single hydrological model, not 20 hydrological models, is used for forecasting. The two forecasting systems have been chosen because systems H and H' differ only in hyper-parameter magnitudes and are both examples of the hyper-parameter tuning approach. The two approaches represent two commonly used hyper-parameter tuning strategies in practice. Hyper-parameter tuning of system H' only involves estimating hyper-parameters from the literature, while hyper-parameter tuning of system H requires more attention and computational costs.

Systems H and H' are henceforth referred as the statistical specific (Ss) system and the statistical uniform (Su) system, respectively, in this paper.

- The Ss system: its hyper-parameters are chosen to optimize forecast performance. As such, the hyper-parameters values here are artificially overestimated and compensate for other sources of uncertainty that are not explicitly accounted for in the EnKF. The optimal hyper-parameters of system Ss are specific to each catchment.
- The Su system: its hyper-parameters describe a more realistic estimate of climate and flow data uncertainties and exhibit more reasonable perturbation magnitudes. The hyper-parameters of system Su are uniform for all catchments.

A brief overview of the Thiboult et al. (2016) hyper-parameter tuning processes, based on processes reported in detail in Thiboult and Anctil (2015), is provided as follows. In both systems, precipitation is perturbed by a Gamma distribution with the mean being the observation

and standard deviation being a proportion of the observation. The proportion has three options (25%, 50%, and 75%). Temperature is perturbed by an additive error that follows a normal distribution with zero mean and two standard deviation options (2℃ and 5℃). Maximum and minimum temperatures are both perturbed by the same additive error random variable. Flow is also perturbed by a normally distributed additive error where the mean error is zero and the standard deviation is proportional to the observed flow at each time step, and the proportion has two options (10% and 25%). The temporal and spatial correlations of errors are not considered in the ensemble generation. All error distributions and their variances constitute the hyper-parameters of the streamflow forecasting experiment. In total, $3 \times 2 \times 2 = 12$ combinations of hyper-parameter values are tested in the hyper-parameter tuning experiments.

Thiboult et al. (2016) also identify the optimal state variables in the process of hyper-parameter tuning to further improve forecast performance. The GR4J model has two potential state variables to be updated in the EnKF. One is the water level of the production store (S), another is the water level of the routing store (R). Updating these states affects the model in different ways, especially regarding the time lag between state updating and the effect on simulated streamflow. Three state combinations (S, R, and both S and R) are tested with the hyper-parameter tuning.

In the hyper-parameter tuning of system Ss, each of the 12 hyper-parameters and state variables are tested. Forecast results are evaluated in terms of reliability and bias in Thiboult and Anctil (2015). Reliability is measured by the Normalized Root-mean-square error Ratio (NRR). Details of NRR are provided in Appendix A.1. Bias is measured by the Nash Sutcliffe efficiency coefficient (NSE) between the observed flow and the forecasted flow ensemble median. The

optimal combination of hyper-parameters and state variables is taken as the one that achieves the best NSE among the three best NRRs. Table 2 summarizes the optimal hyper-parameters and state variables of 20 catchments of system Ss.

Table 2. Optimal hyper-parameters and state variables of 20 catchments of the statistical specific Ss system (from Thiboult et al. (2016)).

Catchment No.	**Precipitation distribution standard deviation proportion**	**Temperature distribution standard deviation (℃)**	**Flow distribution standard deviation proportion**	**State variables**
1	0.75	2	0.1	S-R*
2	0.75	2	0.1	S-R
3	0.25	5	0.1	R
4	0.75	2	0.1	R
5	0.75	2	0.1	S-R
6	0.75	2	0.1	R
7	0.75	2	0.1	S-R
8	0.75	5	0.1	R
9	0.75	2	0.1	S-R
10	0.75	2	0.1	S-R
11	0.75	5	0.1	S-R
12	0.5	2	0.1	S-R
13	0.75	2	0.1	S-R
14	0.5	2	0.1	S-R
15	0.75	2	0.25	S-R
16	0.75	2	0.1	S-R
17	0.75	2	0.1	R
18	0.75	2	0.25	S-R
19	0.75	2	0.1	S-R
20	0.75	2	0.1	S-R

* S-R refers to both S and R.

Here it is worth clarifying the workload of tuning hyper-parameters for system Ss. As mentioned before, there are 12 combinations of hyper-parameters, and three possible choices for the state variables (R, S, and both R and S). This means that there are $12 \times 3 = 36$ combinations for each catchment to be tested. Since we worked on 20 catchments, this requires $36 \times 20 = 720$ ensemble flow forecasting experiments. This is still tolerable as we work with GR4J which only

has two state variables, but in the case where one uses a hydrological model with more state variables, the number of flow forecasting experiments increases dramatically as the number of combinations per catchment model is given by $2^r - 1$, where r is the number of state variables.

The hyper-parameters of system Su are a simpler version of the hyper-parameters of system Ss and are used to describe climate uncertainty more realistically (Thiboult et al., 2016). The standard deviation proportion of precipitation distribution is 25%, the standard deviation of temperature distribution is 2℃, the standard deviation proportion of flow distribution is 10%. The state variables S and R are both updated for every catchment. The hyper-parameter magnitudes of Su are lower than or equal to those of Ss given in Table 2. The climate ensembles generated by the tuned hyper-parameters of Ss and Su are called the traditional ensemble to distinguish from the N15 generated climate ensemble.

2.4.2. Newman et al. (2015) Ensemble Generation

This section presents how to use Newman et al. (2015) dataset, N15, to generate the precipitation and temperature ensemble. N15 has 100 historical realizations of daily total precipitation, mean temperature, and daily temperature range for the period 1980-2012. The dataset covers the contiguous United States, northern Mexico, and southern Canada at $1/8^{\circ}$ resolution. It is free for download at https://www.earthsystemgrid.org/dataset/gridded_precip_and_temp.html.

Newman et al. (2015) Dataset (N15)

N15 is generated from an observation based probabilistic interpolation system. An overview of the probabilistic interpolation system is as follows. The probabilistic interpolation system is

composed of two steps: spatial interpolation and ensemble generation. In spatial interpolation, the probability of precipitation at each grid cell is estimated by a locally weighted logistic regression, and the magnitudes of precipitation and temperature are estimated by a locally weighted linear regression. Both regressions use the latitude, longitude and elevation of neighboring stations as explanatory variables. The climate ensemble is then generated by using the spatially correlated random fields (SCRF) sampled from the standard normal distribution. The SCRF's spatial and temporal correlations are accounted for by the nested grid approach and the autoregressive model of order one, respectively (Fang and Tacher, 2003; Newman et al., 2015). Due to the spatio-temporal correlation and the cross correlation between precipitation and temperature variables in the SCRF generation, each climate ensemble member calculated from the SCRF cannot be substituted or combined with other members across time and space.

N15 has many advantages over the hyper-parameters derived climate ensemble. First, it produces realistic precipitation occurrence by using zero-to-one probability, not zero-or-one probability, to quantify the probability of precipitation, even for zero precipitation observations. Second, it considers the spatial and temporal correlations of the random fields of each grid cell. Third, the high-resolution grid of the dataset enables applications to both lumped and distributed hydrological models. Fourth, the dataset covers an extensive area and thus can be directly used for catchments within the contiguous United States, northern Mexico, and southern Canada. Modelers do not need to repeat the time-consuming error model specification and hyper-parameter tuning processes case by case.

Bias Correction

It is always necessary to check the quality of an ensemble climate dataset before utilizing it in operation because it is found that the meteorological ensembles can be inconsistent with observations, and hydrologically important variables need to be adjusted to be realistic before being used (Graham et al., 2007; Hay et al., 2000; Lenderink et al., 2007; Piani et al., 2010; Yang et al., 2010). This study corrected the bias of precipitation, mean temperature, and daily temperature range at the catchment scale. The detailed bias correction procedure applied here is described in the following paragraph.

Bias refers to the difference between the ensemble mean and the true value of the variable being evaluated. Since the truth is unknown, the observation is taken as the truth here. Taking precipitation as an example, we first calculated the lumped precipitation of each catchment based on the N15 ensemble mean and the observation, respectively. Then we computed the deviations between the catchment ensemble mean and the observed precipitation over time. Based on the above analysis, we found that the precipitation ensemble has a systematic bias in contrast with the observation, while the temperature ensembles do not show systematic bias. The precipitation bias correction was carried out by shifting all the ensemble members at the catchment scale by the same magnitude that equals to the difference between the ensemble mean and the observation at each time step. Bias correction is applied independently to each time step, so the correction magnitude varies with time. In addition, the temperature ensemble members were shifted one day ahead due to the time lag of the ensemble versus the observation (this correction was deemed appropriate based on personal communications with the N15 dataset developer). With these settings, 100 bias corrected realizations of daily precipitation, and 100 corrected realizations of

mean temperature and daily temperature range were defined as the 100-member bias corrected N15 ensemble.

In Thiboult et al. (2016), the EnKF ensemble size is 50, while N15 has 100 members. To provide a fair comparison of the systems, 50 members are randomly selected from the 100-member bias corrected N15 ensemble without replacement. The selected 50 members are referred to as the ensemble N and the corresponding forecasting system is called system N. Except the climate ensemble, system N uses the identical flow perturbations and state variables with system Su in the EnKF phase. Table 3 summaries the EnKF settings of the three systems Su, Ss, and N regarding data perturbations and state variables.

Table 3. EnKF data perturbations and state variables of systems Su, Ss. N. $\tilde{P}$, $\tilde{T}$, and $\tilde{Q}$ represent the measured precipitation, temperature, and flow, respectively. c_P and c_Q are the standard deviation proportions of the precipitation and flow distributions, respectively. σ is the standard deviation of the temperature distribution. Here system Ss contains all the optimum hyper-parameters and state variables of 20 catchments. See Table 2 for details of each catchment.

System	Statistical uniform (Su)	Statistical specific (Ss)	N
Precipitation perturbation	$\mathrm{P}\sim\mathrm{Gamma}(\frac{1}{{c_P}^2}, {c_P}^2 \cdot \tilde{P})$		Derived from N15
	$c_P = 0.25$	$c_P = 0.25$, 0.5 or 0.75	
Temperature perturbation	$\mathrm{T}\sim\mathrm{Normal}(\tilde{T}, \sigma^2)$		
	$\sigma = 2$	$\sigma = 2$ or 5	
Flow perturbation	$\mathrm{Q}\sim\mathrm{Normal}(\tilde{Q}, {c_Q}^2 \cdot \tilde{Q}^2)$		
	$c_Q = 0.1$	$c_Q = 0.1$ or 0.25	$c_Q = 0.1$
State variables	S-R*	S, R, or S-R	S-R

* S-R refers to both S and R.

2.5. Evaluation of Ensemble Flow Forecasts

The flow forecasts of each forecasting system are assessed from two perspectives: deterministic and probabilistic. The metrics below are used to measure flow forecast performances for systems Su, Ss, and N and are selected to be consistent with the metrics used in Thiboult et al. (2016).

Deterministic Forecast

When a deterministic flow forecast is required, the mean of the forecast ensemble is evaluated. The RMSE is used as the deterministic forecast evaluation metric.

$$RMSE = \sqrt{\frac{1}{T}\sum_{t=1}^{T}(\bar{\hat{z}}_t - z_t)^2} \tag{9}$$

where T is the number of time steps of the evaluation period ($t = 1,...,T$). $\bar{\hat{z}}_t$ and z_t are the forecasted ensemble mean flow and the observed flow at time t, respectively. The RMSE is non-negative with the optimum of zero. Units of RMSE are the units of the flows.

Probabilistic Forecast

When probabilistic flow forecasts are required, the ensemble flow forecast results are evaluated by comparing the ensemble flow forecasts with the measured flows. Being consistent with Thiboult et al. (2016), the mean continuous ranked probability score (MCRPS) is adopted as the ensemble forecast evaluation metric. The continuous ranked probability score (CRPS) measures the proximity of the forecast distribution and the measurement distribution at a single time step. When the measurement is deterministic, the CRPS is calculated as (Gneiting and Raftery, 2007):

$$CRPS\left(F_t(\hat{z}), z_t\right) = \int_{-\infty}^{+\infty}\left(F_t(\hat{z}) - H\{\hat{z} \geq z_t\}\right)^2 d\hat{z} \tag{10}$$

where F_t is the cumulative distribution function of ensemble flow forecast at time t. $\hat{z}$ is the forecasted flow. $H\{\hat{z} \geq z_t\}$ is the Heaviside function expressed as:

$$H\{\hat{z} \geq z_t\} = \begin{cases} 1, for\ \hat{z} \geq z_t \\ 0, for\ \hat{z} < z_t \end{cases} \quad (11)$$

The range of CRPS is non-negative with the best value of 0. Units of CRPS are the units of the flows.

MCRPS is the average CRPS over the entire evaluation period. MCRPS is also non-negative with the optimum of zero.

$$MCRPS = \frac{1}{T}\sum_{t=1}^{T} CRPS\left(F_t(\hat{z}), z_t\right) \quad (12)$$

In addition, reliability and spread are assessed for probabilistic flow forecasts. As in Thiboult et al. (2016), the mean absolute error of the reliability diagram (MaeRD) is used to estimate reliability. The reliability diagram is a graph of the observed frequency of an event plotted against the forecast probability of an event (Hartmann and Pagano, 2002). In theory, a perfect forecast system will result in forecasts with a probability of X% being consistent with the observations X% of the time. Hence when plotting a reliability diagram, comparisons are made against the diagonal. A curve above the diagonal line denotes an over-dispersion, an under-dispersion is in the opposite case (Thiboult et al., 2016).

The MaeRD measures the average distance between the forecast probability and the observation probability over all quantiles of interest (Thiboult et al., 2016).

$$MaeRD = \frac{1}{K}\sum_{k=1}^{K}|P_{fcst,k} - P_{obs,k}| \quad (13)$$

where K is the number of quantiles of interest ($k = 1,...,K$). $P_{fcst,k}$ and $P_{obs,k}$ are the forecast probability and observed probability at the k^{th} quantile of interest, respectively. K equals to nine in this study. The MaeRD is dimensionless and non-negative with the optimal value of zero.

Spread is an indicator that should be considered along with reliability because a perfectly reliable forecast at the cost of excessively high dispersion is not desired. The spread equals to the square root of the average ensemble variance over the evaluation period (Fortin et al., 2014).

$$spread = \sqrt{\frac{1}{T}\sum_{t=1}^{T} Var(\hat{z}_t)} \quad (14)$$

$$Var(\hat{z}_t) = \frac{1}{N-1}\sum_{n=1}^{N}(\hat{z}_{t,n} - z_t)^2 \quad (15)$$

where N is the forecast ensemble size ($n = 1,...,N$). $\hat{z}_{t,n}$ is the n^{th} forecasted flow at time t. The spread is non-negative with an optimum of zero and has units of the flows.

3. Results Analysis and Discussion

The comparison results are presented in two sections. Section 3.1 compares the climate ensembles of systems Su, Ss, and N. Section 3.2 compares the flow forecasts of systems Su, Ss, and N.

3.1. Climate Ensemble Comparison

Taking precipitation as an example, Figure 1 compares the 50-member climate ensembles of systems Su, Ss, and N for the Aux Ecorces catchment (the 13th catchment of Table 1) in the

period of July 15-31, 2009. The center line in each box represents the median of the ensemble (q_2), the top and bottom edges of the box are the first and third quartiles of the ensemble (q_1 and q_3). The whiskers above and below the box are the maximum and minimum of the ensemble excluding the outliers. Members are considered as outliers if their value is either greater than $q_3 + w(q_3 - q_1)$ or less than $q_1 - w(q_3 - q_1)$. w equals to 2.7 times of the ensemble standard deviation and corresponds to a 99.3% percent coverage for normally distributed data. Outliers are represented by the small cross outside the whiskers.

Figure 1 demonstrates that the three climate ensembles follow different distributions, and there is a stronger temporal correlation in the N ensemble than in other two ensembles, especially for the measured low and no precipitation events near high ones. For instance, a heavy rain event occurs on July 26, system N is the only system to observe a significant probability of rainfall for the preceding and following days (e.g., July 25, 27). Considering the temporal correlation, it is likely that the traditional ensembles miss parts of this rainfall event.

Figure 1. Here.

In fact, there are a large number of non-zero precipitation estimates in ensemble N when the precipitations are zero in all members of the Su and Ss ensembles. For the entire forecasting period (761 days), a total of $761 \times 2 \times 20 = 30{,}440$ daily climate ensembles are generated for systems Su and Ss over 20 catchments. Among them, 5293 are such zero precipitation ensembles and account for 17.4% of the total ensemble. However, in system N, 94% of the corresponding

5293 N ensembles have at least one ensemble member with a non-zero precipitation. This demonstrates that ensemble N generates a non-zero probability of precipitation when ensembles Su and Ss estimate no precipitation (zero probability). As explained in Section 2.4.2, this capacity is due to the use of zero-to-one probability, not zero-or-one probability, to quantify the probability of precipitation, even for days with precipitation observations equal to zero. This is an appreciable feature of the N ensemble and is beneficial to generate more reliable predictions. Studies have shown that forcing data uncertainty dominate model errors (Carpenter et al., 2001; Slater and Clark, 2006). Climate is the most important driving factor for hydrological models to generate a spread in the state variables, and this spread is essential to map the state variable space in the EnKF (Reichle et al., 2002). With ensembles Su and Ss, when no precipitation is observed, the climate spread is zero. Therefore, the state variable spread would tend to collapse. However, ensemble N preserves a greater state variable spread, even during meteorological periods where precipitation is less likely to occur.

To investigate the overall ensemble spread cross the forecasting period, we compared climate ensemble N with climate ensembles Su and Ss, respectively, in terms of the metric spread. Recall that spread is calculated as the square root of the mean ensemble variance over the evaluation period (Equations (14)-(15)). Based on an analysis of spread across the 50 forecast climate members and 20 catchments for the entire forecasting period, and so a sample size of 50×20=1000, the precipitation spread of ensemble N is always greater than that of Su but smaller than that of Ss for 95% of the 1000 ensembles, the temperature spread of ensemble N is smaller than the spreads of Su and Ss for all the 1000 ensembles.

3.2. Flow Forecast Comparison

RMSE and MCRPS

Figure 2 shows the RMSE and MCRPS of the forecasts issued by systems Su, Ss, and N for 20 catchments. The RMSE and MCRPS are used to demonstrate the deterministic and probabilistic forecast performances, respectively. For brevity, the 2nd, 4th, 5th, and 7th lead day results are not shown. The center of each radar plot corresponds to the optimal metric value, while the outer circle indicates the worse metric value for a given lead time. In Figure 2, system N yields improved or similar results relative to systems Su and Ss over all lead times. This is partly explained by the ability of system N to account for the uncertainty of low and no rainfall events, as mentioned in Section 3.1.

Moreover, the improvements of system N are notable for the 1st and 3rd lead days and diminish with increasing forecast horizons. This phenomenon indicates that the N15 generated climate ensemble is beneficial for improving short-term flow forecast. On the other hand, the decreasing relative improvement as the forecast horizon increases is understandable because data assimilation is known for its dominant influence on shorter horizons, while meteorological ensemble forecasts typically dominate longer ones (Thiboult et al. 2016).

Figure 2. Here.

Reliability and Spread

Figure 3 details the reliability diagrams of systems Su, Ss, and N. All three systems are under-dispersed and are therefore over confident for all lead times. Possible reasons include inaccurate

or biased meteorological forecasts or poorly calibrated model parameters. Another potential reason for only systems Su and N is the lack of full consideration of model errors in the EnKF, so the state variable space is not fully explored. To achieve reliability with a given forecast system, a post-processing of the meteorological and/or hydrological forecasts would be necessary. Some operational guidance can be found in Abaza et al, (2017), Boucher et al. (2015), Raftery et al. (2005), and Rana et al. (2014). This study did not conduct post-processing because post-processing would complicate the comparison of the effects of climate ensemble on flow forecasting.

Figure 3. Here.

Larger differences among the systems are observed from the MaeRD and spread results as depicted in Figure 4. Compared with system Su, system N successfully reduces the MaeRD for more than 70% of the 20 catchments without significant spread changes. Thus, ensemble N produces more reliable flow forecasts than ensemble Su. Compared with system Ss, system N globally worsens the MaeRD. Nonetheless, this decrease of performance needs to be qualified since the reliability of system Ss is achieved through inflated hyper-parameters in order to get more reliable hydrological ensembles by indirectly accounting for additional sources of uncertainty. Like in Figure 2, the metric differences between all three systems diminish with increasing forecast horizons as the data assimilation dominates short-term forecasts (Thiboult et al. 2016).

Figure 4. Here.

It is not unexpected that system Ss gets wider prediction spread than system N because system Ss uses higher and catchment specific perturbation magnitudes to account for other model errors in the EnKF. High climate uncertainty can be propagated to model outputs generating wide prediction intervals that contain more flow observations. From this point of view, the under-dispersion problem of system N can be fixed by taking into account other model errors that are specific to the catchment and the hydrological model in the EnKF.

In terms of incorporating other model errors in the EnKF, one can consider the errors of other climate inputs, model parameter and model structure. For example, Reichle et al. (2002) add errors to wind speed, short- and long-wave radiative flux, and surface pressure in addition to precipitation and temperature. Clark et al. (2006) consider the model parameter uncertainty in the EnKF by generating 100 parameter sets corresponding to 100 climate ensemble members via the Monte Carlo Markov Chains.

Hydrograph

To visualize the flow forecast improvements due to N15, Figure 5 illustrates the first lead day flow forecasts of systems Su, Ss, and N for the Aux Ecorces catchment (the 13th catchment of Table 1) and a portion of the forecasting period. The ensemble mean and the 95% prediction intervals of forecasted flows are shown. Figure 5 confirms the forecast performance upgrade by using ensemble N. The deterministic flow forecasts are more consistent with the observations in system N than in systems Su and Ss. Also, more observations fall into the 95% prediction

envelope of system N than that of systems Su and Ss. Specifically, looking at the peak flow forecasts in May 2009, the ensemble means of systems Su and Ss always underestimate the peaks. The 95% prediction intervals of systems Su and Ss cover only a small part of the observations. In comparison, system N improves the consistency between the ensemble mean and the peak, and its 95% forecast intervals contain almost all the peak flows.

Figure 5. Here.

4. Conclusions and Future Work

This paper proposed using an existing ensemble climate product to efficiently represent climate data uncertainty in the EnKF of short-term ensemble flow forecasting. Specifically, this study used the Newman et al. (2015) dataset, referred to here as N15, to represent measured precipitation and temperature uncertainties. To our knowledge, this is the first study to compare the N15 generated climate ensemble with the carefully tuned hyper-parameter generated climate ensemble in real ensemble flow forecasting. The tuned hyper-parameters are from two ensemble forecasting systems developed in Thiboult et al. (2016). One is system Su, a subset of system H' in Thiboult et al. (2016), that uses somewhat realistic perturbation magnitudes to represent climate uncertainty. Another is system Ss, a subset of system H in Thiboult et al. (2016), that uses unrealistically inflated hyper-parameters to implicitly compensate unaccounted model errors and optimize forecast performance. The hyper-parameters of the two systems are carefully tuned by Thiboult and Anctil (2015). Another highlight of this study is that a large number of experiments (20 Québec catchments) are explored to draw general conclusions.

The climate ensemble comparison shows that there is a stronger temporal correlation in the N15 generated climate ensemble than in the traditional ensembles Su and Ss, especially for the measured low and no precipitation events near high precipitations. N15 estimates a non-zero probability of precipitation when both traditional ensembles predict zero probability of precipitation. The ensemble flow forecast comparison of 20 catchments demonstrates that the N15 generated ensemble yields improved or similar deterministic and probabilistic flow forecasts relative to both traditional climate ensembles as measured by a variety of performance metrics (i.e., RMSE, MCRP, reliability diagram, MaeRD, and spread). The relative improvement of N15 derived forecasts is especially significant for short lead times (i.e., 1-3 days in our case) when the influence of the data assimilation dominates.

Our comparison study suggests that it is possible to eliminate the need for precipitation and temperature relevant hyper-parameter tuning from the EnKF by using an example historical ensemble climate product without losing flow forecast performance. Moreover, two gains are: (1) saving a great amount of time and computational cost from hyper-parameter tuning, and (2) partly disaggregating sources of uncertainty in the EnKF by explicitly addressing precipitation and temperature uncertainties. The saved efforts can be used to incorporate unaccounted model errors (e.g., model parameter and structure errors). Explicit consideration of uncertainty sources in the EnKF is critical in the pursuit of the right results for the right reasons.

N15 is an effective resource of historical climate ensemble for the contiguous United States, northern Mexico, and southern Canada. This resource provides realistic and spatio-temporally correlated precipitation and temperature uncertainty estimates. It is straightforward to be used for

both lumped and distributed hydrological model applications. Moreover, its climate ensemble generation method is applicable to real-time flow forecasting as its real-time ensemble generation code is available (readers can contact the N15 dataset developers to get the code). Future work should investigate the use of other historical climate ensemble products according to research area and data availability.

There are two areas for improving our research. First, our findings depend on the fact that we considered the initial condition uncertainty and the input and output data uncertainty in the EnKF. The forecasting performance improvements may not be as significant as they are presented here after adding other model errors (e.g., model structure and parameter errors) due to the compensation effect. In the future, more experiments are needed to explicitly incorporate the model parameter and model structure uncertainties in the EnKF. This could be achieved by, for example, using multiple parameter sets per model and multiple hydrological models in the state update, respectively. After taking into account other uncertainties, the authors will explore if the N15 based forecast broadens the prediction intervals without deteriorating the overall forecast performance. The second future research area is to investigate the transferability of the current findings to other hydrological models. Our research is built on a single hydrological model (i.e., GR4J), and though the results are promising, the N15 based forecasting needs to be validated for a variety of hydrological models, in particular, distributed hydrological models.

Acknowledgement

This work was supported by the Natural Science and Engineering Research Council (NSERC) Canadian FloodNet (Grant number: NETGP 451456). We appreciate the support of Dr. Andrew Newman for his help in utilizing the Newman et al. (2015) dataset.

References

Abaza, M., Anctil, F., Fortin, V., Perreault, L., 2017. On the incidence of meteorological and hydrological processors: Effect of resolution, sharpness and reliability of hydrological ensemble forecasts. J. Hydrol. 555, 371–384. doi:10.1016/j.jhydrol.2017.10.038

Abaza, M., Anctil, F., Fortin, V., Turcotte, R., 2014a. Sequential streamflow assimilation for short-term hydrological ensemble forecasting. J. Hydrol. 519, 2692–2706. doi:10.1016/J.JHYDROL.2014.08.038

Abaza, M., Garneau, C., Anctil, F., 2014b. Comparison of Sequential and Variational Streamflow Assimilation Techniques for Short-Term Hydrological Forecasting. J. Hydrol. Eng. 20, 04014042. doi:10.1061/(ASCE)HE.1943-5584.0001013

Bergeron, O., 2016. Guide d'utilisation 2016 - Grilles climatiques quotidiennes du Programme de surveillance du climat du Québec (Version 1.2). Québec.

Boucher, M.-A., Perreault, L., Anctil, F., Favre, A.-C., 2015. Exploratory analysis of statistical post-processing methods for hydrological ensemble forecasts. Hydrol. Process. 29, 1141–1155. doi:10.1002/hyp.10234

Carpenter, T.M., Georgakakos, K.P., Sperfslagea, J.A., 2001. On the parametric and NEXRAD-radar sensitivities of a distributed hydrologic model suitable for operational use. J. Hydrol. 253, 169–193. doi:10.1016/S0022-1694(01)00476-0

Clark, M.P., Rupp, D.E., Woods, R.A., Zheng, X., Ibbitt, R.P., Slater, A.G., Schmidt, J.,

Uddstrom, M.J., 2008. Hydrological data assimilation with the ensemble Kalman filter: Use of streamflow observations to update states in a distributed hydrological model. Adv. Water Resour. 31, 1309–1324. doi:10.1016/j.advwatres.2008.06.005

Clark, M.P., Slater, A.G., 2006. Probabilistic quantitative precipitation estimation in complex terrain. J. Hydrometeorol. 7, 3–22. doi:10.1175/JHM474.1

Clark, M.P., Slater, A.G., Barrett, A.P., Hay, L.E., McCabe, G.J., Rajagopalan, B., Leavesley, G.H., 2006. Assimilation of snow covered area information into hydrologic and land-surface models. Adv. Water Resour. 29, 1209–1221. doi:10.1016/j.advwatres.2005.10.001

Del Giudice, D., Albert, C., Rieckermann, J., Reichert, P., 2016. Describing the catchment-averaged precipitation as a stochastic process improves parameter and input estimation. Water Resour. Res. 52, 3162–3186. doi:10.1002/2015WR017871

Del Giudice, D., Löwe, R., Madsen, H., Mikkelsen, P.S., Rieckermann, J., 2015. Comparison of two stochastic techniques for reliable urban runoff prediction by modeling systematic errors. Water Resour. Res. 51, 5004–5022. doi:10.1002/2014WR016678

Demirel, M.C., Booij, M.J., Hoekstra, A.Y., 2013. Effect of different uncertainty sources on the skill of 10 day ensemble low flow forecasts for two hydrological models. Water Resour. Res. 49, 4035–4053. doi:10.1002/wrcr.20294

Duan, Q., Sorooshian, S., Gupta, V., 1992. Effective and efficient global optimization for conceptual rainfall-runoff models. Water Resour. Res. 28, 1015–1031. doi:10.1029/91WR02985

Dunne, S., Entekhabi, D., 2006. Land surface state and flux estimation using the ensemble Kalman smoother during the Southern Great Plains 1997 field experiment. Water Resour. Res. 42. doi:10.1029/2005WR004334

Eicker, A., Schumacher, M., Kusche, J., Döll, P., Schmied, H.M., 2014. Calibration/Data Assimilation Approach for Integrating GRACE Data into the WaterGAP Global Hydrology Model (WGHM) Using an Ensemble Kalman Filter: First Results. Surv. Geophys. 35, 1285–1309. doi:10.1007/s10712-014-9309-8

Evensen, G., 2007. Data Assimilation, The Ensemble Kalman Filter, Springer. Springer Berlin Heidelberg, Berlin, Heidelberg. doi:10.1007/978-3-540-38301-7

Evensen, G., 2003. The Ensemble Kalman Filter: Theoretical formulation and practical implementation. Ocean Dyn. 53, 343–367. doi:10.1007/s10236-003-0036-9

Evensen, G., 1994. Sequential data assimilation with a nonlinear quasi-geostrophic model using Monte Carlo methods to forecast error statistics. J. Geophys. Res. 99, 10143. doi:10.1029/94JC00572

Fang, J., Tacher, L., 2003. An efficient and accurate algorithm for generating spatially-correlated random fields. Commun. Numer. Methods Eng. 19, 801–808. doi:10.1002/cnm.621

Fortin, V., Abaza, M., Anctil, F., Turcotte, R., Fortin, V., Abaza, M., Anctil, F., Turcotte, R., 2014. Why Should Ensemble Spread Match the RMSE of the Ensemble Mean? J. Hydrometeorol. 15, 1708–1713. doi:10.1175/JHM-D-14-0008.1

Gneiting, T., Raftery, A.E., 2007. Strictly Proper Scoring Rules, Prediction, and Estimation. J. Am. Stat. Assoc. 102, 359–378. doi:10.1198/016214506000001437

Graham, L.P., Hagemann, S., Jaun, S., Beniston, M., 2007. On interpreting hydrological change from regional climate models. Clim. Change 81, 97–122. doi:10.1007/s10584-006-9217-0

Hamill, T.M., Snyder, C., 2000. A Hybrid Ensemble Kalman Filter–3D Variational Analysis Scheme. Mon. Weather Rev. 128, 2905–2919. doi:10.1175/1520-0493(2000)128<2905:AHEKFV>2.0.CO;2

Hartmann, H., Pagano, T., 2002. Confidence builders: Evaluating seasonal climate forecasts from user perspectives. Bull.

Hay, L.E., Wilby, R.L., Leavesley, G.H., 2000. A comparison of delta change and downscaled GCM scenarios for three mountainous basins in the United States. J. Am. Water Resour. Assoc. 36, 387–397. doi:10.1111/j.1752-1688.2000.tb04276.x

Heemink, A.W., Verlaan, M., Segers, A.J., 2001. Variance Reduced Ensemble Kalman Filtering. Mon. Weather Rev. 129, 1718–1728. doi:10.1175/1520-0493(2001)129<1718:VREKF>2.0.CO;2

Houtekamer, P.L., Mitchell, H.L., 2001. A Sequential Ensemble Kalman Filter for Atmospheric Data Assimilation. Mon. Weather Rev. 129, 123–137. doi:10.1175/1520-0493(2001)129<0123:ASEKFF>2.0.CO;2

Houtekamer, P.L., Mitchell, H.L., 1998. Data Assimilation Using an Ensemble Kalman Filter Technique. Mon. Weather Rev. 126, 796–811. doi:10.1175/1520-0493(1998)126<0796:DAUAEK>2.0.CO;2

Huang, C., Newman, A.J., Clark, M.P., Wood, A.W., Zheng, X., 2017. Evaluation of snow data assimilation using the ensemble Kalman filter for seasonal streamflow prediction in the western United States. Hydrol. Earth Syst. Sci. 21, 635–650. doi:10.5194/hess-21-635-2017

Kavetski, D., Kuczera, G., Franks, S.W., 2006a. Bayesian analysis of input uncertainty in hydrological modeling: 1. Theory. Water Resour. Res. 42, W03407. doi:10.1029/2005WR004368

Kavetski, D., Kuczera, G., Franks, S.W., 2006b. Bayesian analysis of input uncertainty in hydrological modeling: 2. Application. Water Resour. Res. 42, W03408. doi:10.1029/2005WR004376

Khaki, M., Schumacher, M., Forootan, E., Kuhn, M., Awange, J.L., van Dijk, A.I.J.M., 2017. Accounting for spatial correlation errors in the assimilation of GRACE into hydrological models through localization. Adv. Water Resour. 108, 99–112. doi:10.1016/j.advwatres.2017.07.024

Kumar, S. V., Dong, J., Peters-Lidard, C.D., Mocko, D., Gómez, B., 2016. Role of forcing uncertainty and model error background characterization in snow data assimilation. Hydrol. Earth Syst. Sci. Discuss. 1–24. doi:10.5194/hess-2016-581

Leisenring, M., Moradkhani, H., 2011. Snow water equivalent prediction using Bayesian data assimilation methods. Stoch. Environ. Res. Risk Assess. 25, 253–270. doi:10.1007/s00477-010-0445-5

Lenderink, G., Buishand, A., van Deursen, W., 2007. Estimates of future discharges of the river Rhine using two scenario methodologies: direct versus delta approach. Hydrol. Earth Syst. Sci. 11, 1145–1159. doi:10.5194/hess-11-1145-2007

Li, Y., Ryu, D., Western, A.W., Wang, Q.J., Robertson, D.E., Crow, W.T., 2014. An integrated error parameter estimation and lag-aware data assimilation scheme for real-time flood forecasting. J. Hydrol. 519, 2722–2736. doi:10.1016/J.JHYDROL.2014.08.009

Liu, Y., Gupta, H.V., 2007. Uncertainty in hydrologic modeling: Toward an integrated data assimilation framework. Water Resour. Res. 43, W07401. doi:10.1029/2006WR005756

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 Resour. Res. 53, 2199–2239. doi:10.1002/2016WR019168

McMillan, H.K., Hreinsson, E.Ö., Clark, M.P., Singh, S.K., Zammit, C., Uddstrom, M.J., 2013.

Operational hydrological data assimilation with the recursive ensemble Kalman filter. Hydrol. Earth Syst. Sci. 17, 21–38. doi:10.5194/hess-17-21-2013

Moradkhani, H., Sorooshian, S., Gupta, H. V., Houser, P.R., 2005. Dual state-parameter estimation of hydrological models using ensemble Kalman filter. Adv. Water Resour. 28, 135–147. doi:10.1016/j.advwatres.2004.09.002

Newman, A.J., Clark, M.P., Craig, J., Nijssen, B., Wood, A., Gutmann, E., Mizukami, N., Brekke, L., Arnold, J.R., 2015. Gridded ensemble precipitation and temperature estimates for the contiguous United States. J. Hydrometeorol. 16, 2481–2500. doi:10.1175/JHM-D-15-0026.1

Noh, S.J., Rakovec, O., Weerts, A.H., Tachikawa, Y., 2014. On noise specification in data assimilation schemes for improved flood forecasting using distributed hydrological models. J. Hydrol. 519, 2707–2721. doi:10.1016/J.JHYDROL.2014.07.049

Oudin, L., Michel, C., Anctil, F., 2005. Which potential evapotranspiration input for a lumped rainfall-runoff model? Part 1 - Can rainfall-runoff models effectively handle detailed potential evapotranspiration inputs? J. Hydrol. 303, 275–289. doi:10.1016/j.jhydrol.2004.08.025

Pauwels, V.R.N., De Lannoy, G.J.M., Pauwels, V.R.N., Lannoy, G.J.M. De, 2006. Improvement of Modeled Soil Wetness Conditions and Turbulent Fluxes through the Assimilation of Observed Discharge. J. Hydrometeorol. 7, 458–477. doi:10.1175/JHM490.1

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

Piani, C., Haerter, J.O., Coppola, E., 2010. Statistical bias correction for daily precipitation in regional climate models over Europe. Theor. Appl. Climatol. 99, 187–192.

doi:10.1007/s00704-009-0134-9

Raftery, A.E., Gneiting, T., Balabdaoui, F., Polakowski, M., 2005. Using Bayesian Model Averaging to Calibrate Forecast Ensembles. Mon. Weather Rev. 133, 1155–1174. doi:10.1175/MWR2906.1

Rana, A., Foster, K., Bosshard, T., Olsson, J., Bengtsson, L., 2014. Impact of climate change on rainfall over Mumbai using Distribution-based Scaling of Global Climate Model projections. J. Hydrol. Reg. Stud. 1, 107–128. doi:10.1016/J.EJRH.2014.06.005

Rasmussen, J., Madsen, H., Jensen, K.H., Refsgaard, J.C., 2015. Data assimilation in integrated hydrological modeling using ensemble Kalman filtering: evaluating the effect of ensemble size and localization on filter performance. Hydrol. Earth Syst. Sci 19, 2999–3013. doi:10.5194/hess-19-2999-2015

Reichle, R.H., Koster, R.D., 2003. Assessing the Impact of Horizontal Error Correlations in Background Fields on Soil Moisture Estimation. J. Hydrometeorol. 4, 1229–1242. doi:10.1175/1525-7541(2003)004<1229:ATIOHE>2.0.CO;2

Reichle, R.H., Walker, J.P., Koster, R.D., Houser, P.R., 2002. Extended versus Ensemble Kalman Filtering for Land Data Assimilation. J. Hydrometeorol. 3, 728–740. doi:10.1175/1525-7541(2002)003<0728:EVEKFF>2.0.CO;2

Renard, B., Kavetski, D., Kuczera, G., Thyer, M., Franks, S.W., 2010a. Understanding predictive uncertainty in hydrologic modeling: The challenge of identifying input and structural errors. Water Resour. Res. 46. doi:10.1029/2009WR008328

Renard, B., Kavetski, D., Kuczera, G., Thyer, M., Franks, S.W., 2010b. Understanding predictive uncertainty in hydrologic modeling: The challenge of identifying input and structural errors. Water Resour. Res. 46, W05521. doi:10.1029/2009WR008328

Renard, B., Kavetski, D., Leblois, E., Thyer, M., Kuczera, G., Franks, S.W., 2011. Toward a reliable decomposition of predictive uncertainty in hydrological modeling: Characterizing rainfall errors using conditional simulation. Water Resour. Res. 47. doi:10.1029/2011WR010643

Salamon, P., Feyen, L., 2010. Disentangling uncertainties in distributed hydrological modeling using multiplicative error models and sequential data assimilation. Water Resour. Res. 46, W12501. doi:10.1029/2009WR009022

Salamon, P., Feyen, L., 2009. Assessing parameter, precipitation, and predictive uncertainty in a distributed hydrological model using sequential data assimilation with the particle filter. J. Hydrol. 376, 428–442. doi:10.1016/j.jhydrol.2009.07.051

Sikorska, A.E., Scheidegger, A., Banasik, K., Rieckermann, J., 2012. Bayesian uncertainty assessment of flood predictions in ungauged urban basins for conceptual rainfall-runoff models. Hydrol. Earth Syst. Sci. 16, 1221–1236. doi:10.5194/hess-16-1221-2012

Slater, A.G., Clark, M.P., 2006. Snow Data Assimilation via an Ensemble Kalman Filter. J. Hydrometeorol. 7, 478–493. doi:10.1175/JHM505.1

Tangdamrongsub, N., Steele-Dunne, S.C., Gunter, B.C., Ditmar, P.G., Weerts, A.H., 2015. Data assimilation of GRACE terrestrial water storage estimates into a regional hydrological model of the Rhine River basin. Hydrol. Earth Syst. Sci. 19, 2079–2100. doi:10.5194/hess-19-2079-2015

Thiboult, A., Anctil, F., 2015. On the difficulty to optimally implement the Ensemble Kalman filter: An experiment based on many hydrological models and catchments. J. Hydrol. 529, 1147–1160. doi:10.1016/J.JHYDROL.2015.09.036

Thiboult, A., Anctil, F., Boucher, M.A., 2016. Accounting for three sources of uncertainty in

ensemble hydrological forecasting. Hydrol. Earth Syst. Sci. 20, 1809–1825. doi:10.5194/hess-20-1809-2016

Tippett, M.K., Anderson, J.L., Bishop, C.H., Hamill, T.M., Whitaker, J.S., 2003. Ensemble Square Root Filters. Mon. Weather Rev. 131, 1485–1490. doi:10.1175/1520-0493(2003)131<1485:ESRF>2.0.CO;2

Valéry, A., Andréassian, V., Perrin, C., 2014. "As simple as possible but not simpler": What is useful in a temperature-based snow-accounting routine? Part 1 - Comparison of six snow accounting routines on 380 catchments. J. Hydrol. 517, 1176–1187. doi:10.1016/j.jhydrol.2014.04.059

Vrugt, J.A., Diks, C.G.H., Gupta, H. V, Bouten, W., Verstraten, J.M., 2005. Improved treatment of uncertainty in hydrologic modeling: Combining the strengths of global optimization and data assimilation. Water Resour. Res. 41, W01017. doi:10.1029/2004WR003059

Wang, S., Huang, G.H., Baetz, B.W., Cai, X.M., Ancell, B.C., Fan, Y.R., 2017. Examining dynamic interactions among experimental factors influencing hydrologic data assimilation with the ensemble Kalman filter. J. Hydrol. 554, 743–757. doi:10.1016/j.jhydrol.2017.09.052

Weerts, A.H., El Serafy, G.Y.H., 2006. Particle filtering and ensemble Kalman filtering for state updating with hydrological conceptual rainfall-runoff models. Water Resour. Res. 42. doi:10.1029/2005WR004093

Whitaker, J.S., Hamill, T.M., 2002. Ensemble Data Assimilation without Perturbed Observations. Mon. Weather Rev. 130, 1913–1924. doi:10.1175/1520-0493(2002)130<1913:EDAWPO>2.0.CO;2

Wright, A.J., Walker, J.P., Pauwels, V.R.N., 2017. Estimating rainfall time series and model

parameter distributions using model data reduction and inversion techniques. Water Resour. Res. 53, 6407–6424. doi:10.1002/2017WR020442

Yang, W., Andréasson, J., Phil Graham, L., Olsson, J., Rosberg, J., Wetterhall, F., 2010. Distribution-based scaling to improve usability of regional climate model projections for hydrological climate change impacts studies. Hydrol. Res. 41, 211–229. doi:10.2166/nh.2010.004

1 Figure caption list

Figure 6. 50-member climate precipitation ensembles of systems Su, Ss, and N for the Aux Ecorces catchment (the 13th catchment of Table 1) and the period of July 15-31, 2009. See Table 3 for the descriptions of systems Su, Ss, and N.

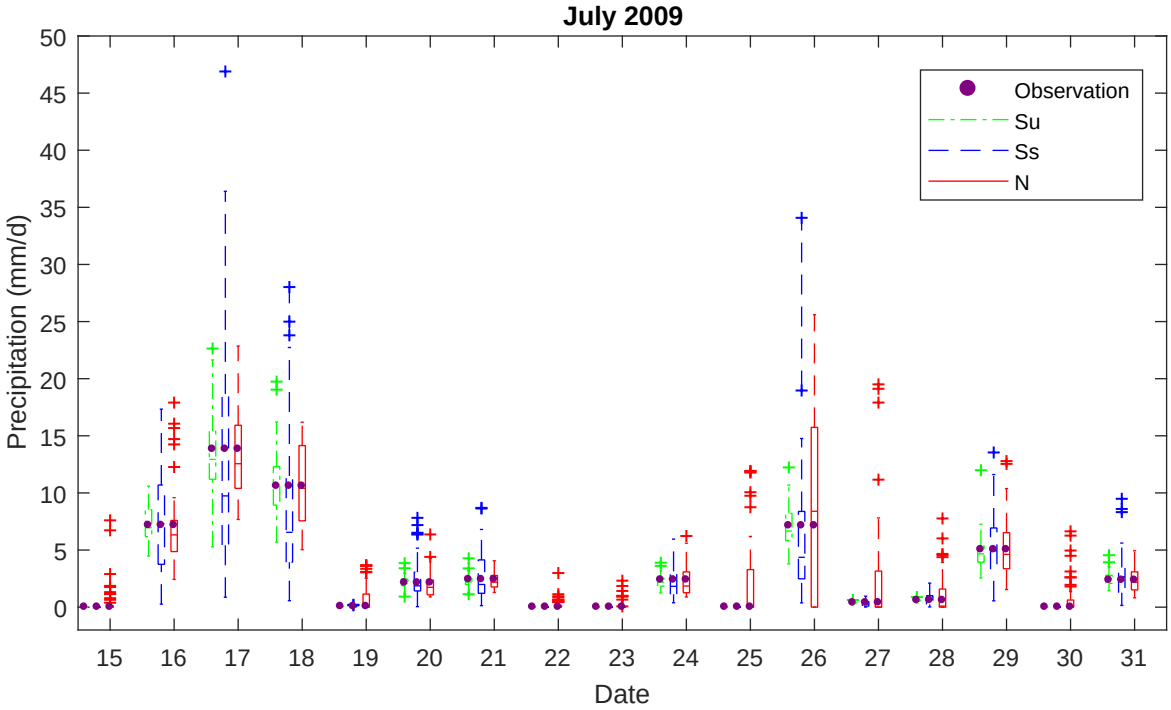
Figure 7. RMSE and MCRPS of systems Su, Ss, and N of 20 catchments for the 1st, 3rd, 6th, and 9th lead days flow forecasts. The RMSE is the root mean square error between the forecasted ensemble mean and the observed flows. The MCRPS is the mean continuous ranked probability score between the forecasted ensemble and the observed flows. See Table 3 for the descriptions of systems Su, Ss, and N. Each catchment is identified by the label on the outer edge of the circle and the catchment metric result is represented by the value on the corresponding spoke. Metric values radiate outward on spokes from the central value of zero (optimal).

Figure 8. Reliability diagrams of systems Su, Ss and N for all 20 catchments and the 1st, 3rd, 6th, and 9th lead days flow forecasts. Each curve of a reliability diagram refers to a catchment. The diagonal line represents the perfectly reliable forecast. See Table 3 for the descriptions of systems Su, Ss, and N.

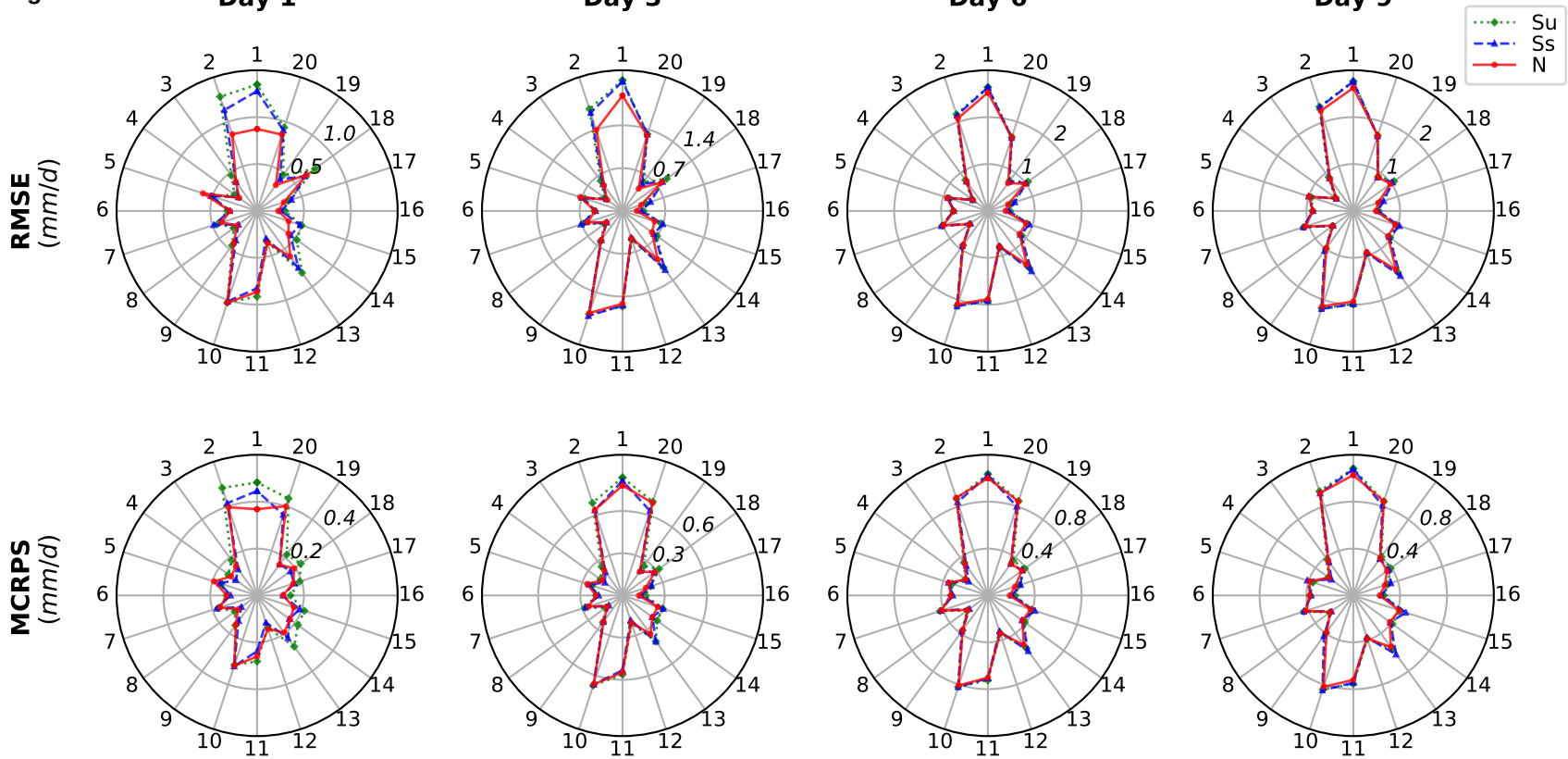
Figure 9. MaeRD and spread of systems Su, Ss and N of 20 catchments for the 1st, 3rd, 6th, and 9th lead days flow forecasts. The MaeRD is the average distance between the forecast probability and the observation probability over nine quantiles. The spread is the square root of the average variance of forecasted flow ensemble. See Table 3 for the descriptions of systems Su, Ss, and N. Each catchment is identified by the label on the outer edge of the circle and the catchment metric result is represented by the value on the corresponding spoke. Metric values radiate outward on spokes from the central value of zero.

Figure 10. Flow forecasts of systems Su, Ss, and N for the Aux Ecorces catchment (the 13th catchment of Table 1) and a portion of the forecasting period. See Table 3 for the descriptions of systems Su, Ss, and N.

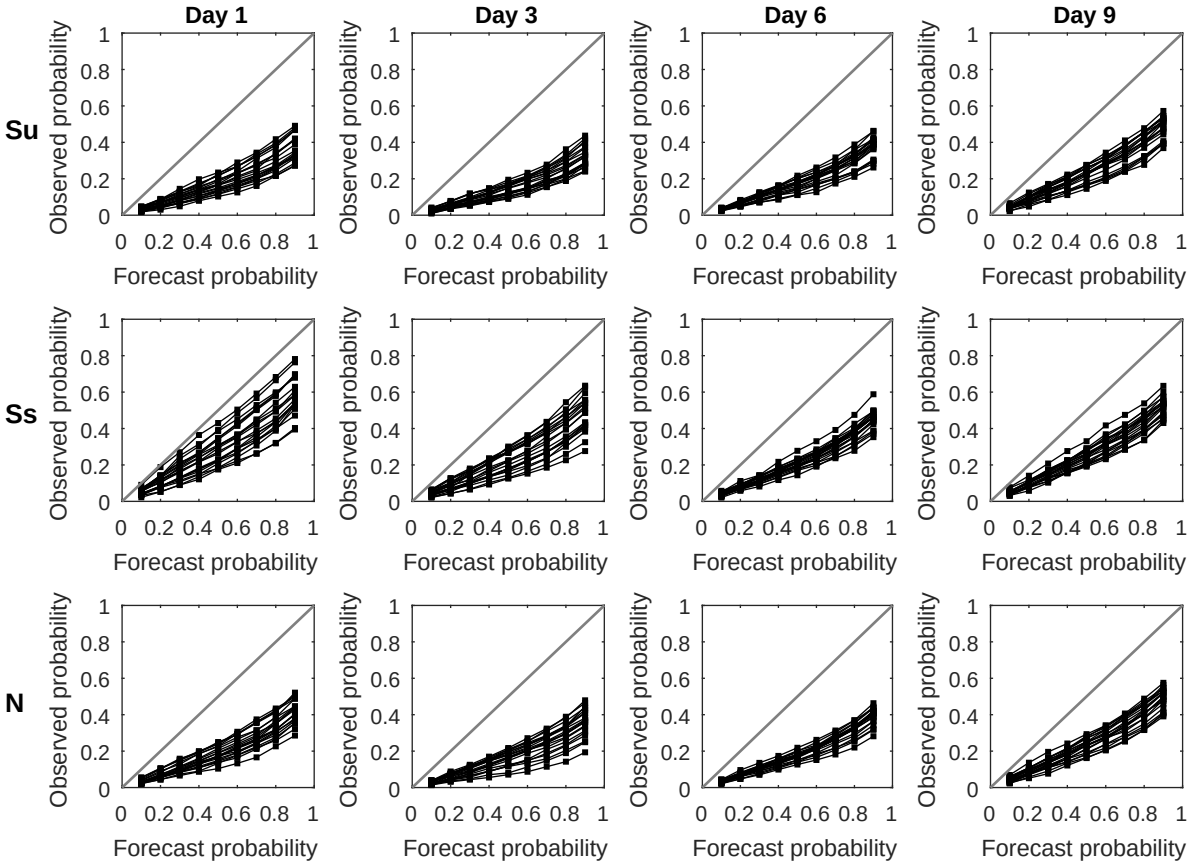
Figure


July 2009
Precipitation (mm/d)
Date
Observation
Su
Ss
N
0
5
10
15
20
25
30
35
40
45
50
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31

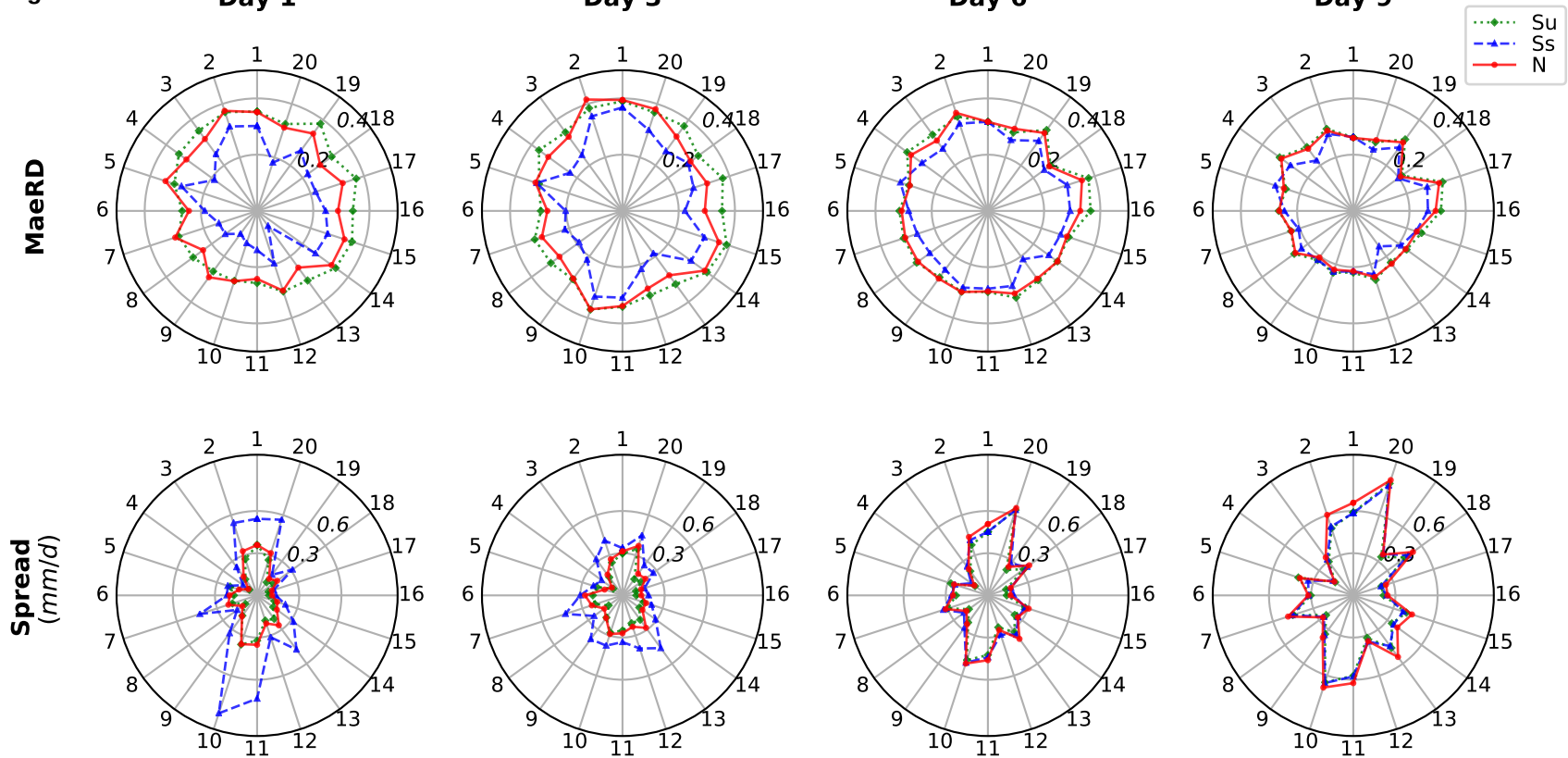
Figure


Day 1
Day 3
Day 6
Day 9
Su
Ss
N
RMSE
(mm/d)
MCRPS
(mm/d)
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
0.5
1.0
0.7
1.4
1
2
0.2
0.4
0.3
0.6
0.4
0.8

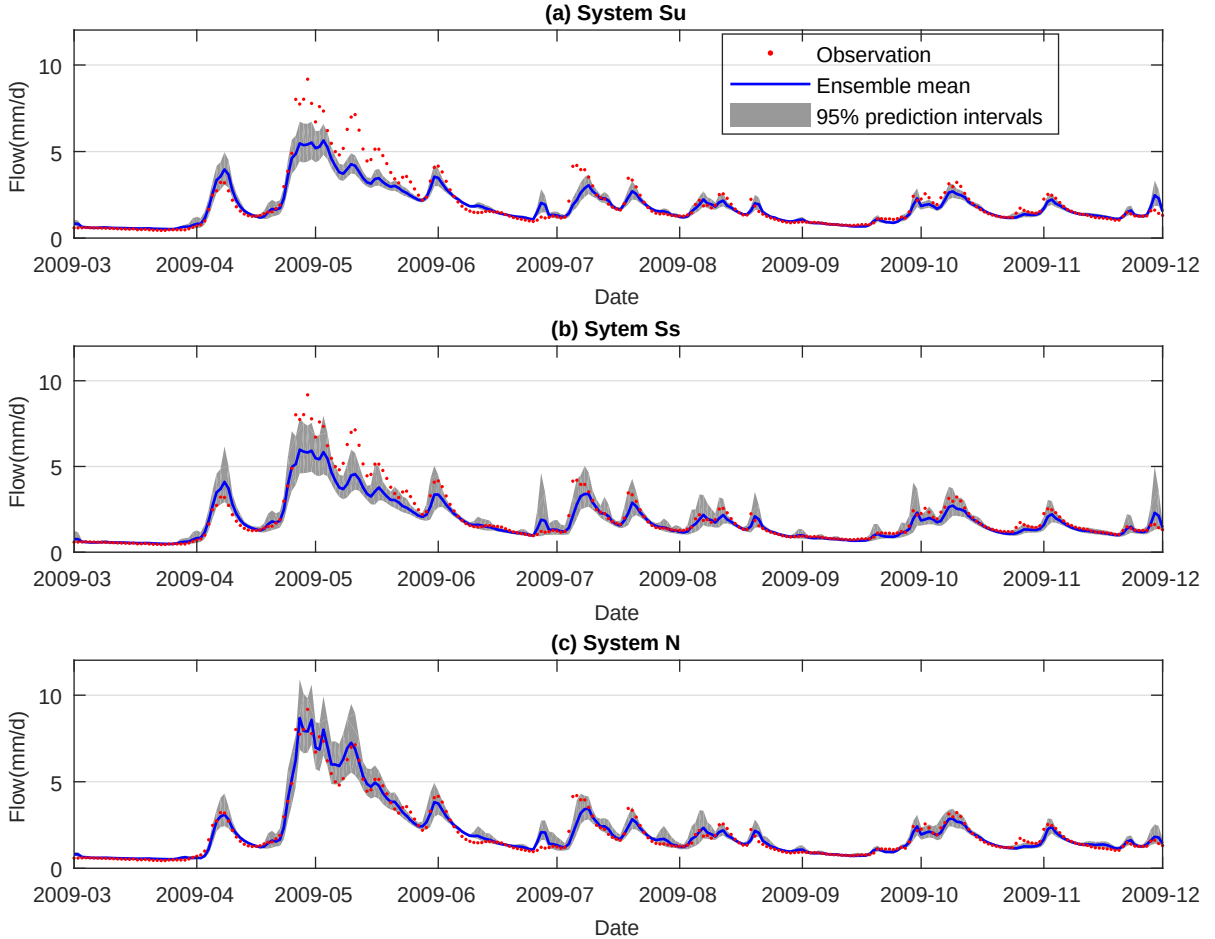
Figure


Day 1
Day 3
Day 6
Day 9
Su
Ss
N
Observed probability
Forecast probability
0
0.2
0.4
0.6
0.8
1

Figure


Day 1
Day 3
Day 6
Day 9
Su
Ss
N
MaeRD
Spread (mm/d)
0.2
0.4
0.3
0.6
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20

Figure



Highlights

- EnKF-based flow forecasts use an existing historical climate ensemble product (N15)
- N15 gives more realistic climate uncertainty than tuned statistical error models
- N15 yields improved or similar flow forecast quality
- N15 saves users from tuning precipitation and temperature relevant hyper-parameters
- N15 enables disaggregating hydrologic model uncertainty sources in the EnKF