

On the decadal predictability of the frequency of flood events across the U.S. Midwest

ANDREA NERI¹, GABRIELE VILLARINI², KAUSTUBH A. SALVI³, LOUISE J. SLATER⁴,
AND FRANCESCO NAPOLITANO¹

¹ Department of Civil, Constructional and Environmental Engineering, University of Rome “La Sapienza”, Rome, Italy.

² IIHR-Hydroscience & Engineering, The University of Iowa, Iowa City, Iowa, USA.

³ Department of Civil Engineering, College of Engineering, Pune, India.

⁴ School of Geography and the Environment, University of Oxford, Oxford, UK.

Corresponding author address:

Andrea Neri, Department of Civil, Constructional and Environmental Engineering, University of Rome “La Sapienza”, Via Eudossiana 18, 00184, Rome, Italy. E-mail: andrea.neri@uniroma1.it.

Tel.: +390644585063

This article has been accepted for publication and undergone full peer review but has not been through the copyediting, typesetting, pagination and proofreading process, which may lead to differences between this version and the Version of Record. Please cite this article as doi: 10.1002/joc.5915

ABSTRACT

Skillful predictions of the frequency of flood events over long lead times (e.g., from one to ten years ahead) are essential for governments and institutions making near-term flood risk plans. However, little is known about current flood prediction capabilities over annual to decadal time scales. Here we address this knowledge gap at 286 U.S. Geological Survey gaging stations across the U.S. Midwest using precipitation and temperature decadal predictions from the Coupled Model Intercomparison Project (CMIP) phase 5 models. We use the 1- to 10-year predictions of precipitation and temperature as inputs to statistical models that have significant skill in reproducing inter-annual and decadal changes in the observed frequency of flood events. Our results indicate that the limited skill of basin-averaged precipitation predictions suppresses the skill of flood event frequency predictions, even at the shortest lead time, but downscaling and bias correction improves predictions across all lead times and especially in spring.

Keywords: flood frequency; decadal prediction; statistical modeling; seasonal; peak-over-threshold; general circulation model

1 Introduction

Flooding is one of the most destructive natural hazards in the United States, claiming many fatalities and significant economic damages each year. Economic losses caused by flood events increased during the last century (Downton and Pielke, 2005), as did the number of fatalities during the last decade of the 20th century (Kunkel et al., 1999). During the years 1959-2005, there were a total of 4,586 reported deaths caused by flooding, with an average of 97.6 fatalities per year (Ashley and Ashley, 2008). Given the magnitude and societal significance of these impacts, the prediction of flood events represents a key step towards reducing damages and choosing the most appropriate flood mitigation plans (Carsell et al., 2004).

The vast majority of published studies related to flood predictions have focused on predictions with short (e.g., days) to medium (e.g., months) lead times, generally using outputs from numerical weather prediction models (NWP) or general circulation models (GCMs) as input to hydrological or statistical models (e.g., Cloke and Pappenberger, 2009; Slater and Villarini, 2018). For instance, He et al. (2010) used outputs from The Observing System Research and Predictability Experiment Interactive Grand Global Ensemble (TIGGE; Bougeault et al., 2010) dataset to predict flood discharge in the Huai River catchment (China) up to 10 days in advance using the Xinanjiang hydrological model, with promising results. Wood and Lettenmaier (2006) proposed a seasonal hydrologic forecasting system which provides ensemble streamflow forecasts over monthly to seasonal lead times in many catchments across the western United States. The European Flood Awareness System (Thiemig and Salamon, 2017) monitors and forecasts floods across Europe for lead times ranging from 1 to 10 days, using the European Centre for Medium-Range Weather Forecasts to force the LISFLOOD (van der Knijff et al., 2010) hydrological model. Alongside the physically-based models, data-driven approaches are

also commonly adopted to predict floods at the monthly/seasonal scale, using initial hydrologic conditions and atmosphere-oceanic indices as predictors of statistical models (Garen, 1993; Mendoza et al., 2017; Piechota et al., 1998; Plummer et al., 2009; Robertson and Wang, 2012).

Seasonal and sub-seasonal forecasts of discharge are essential components of flood warning systems, developed by many organizations including THORPEX (The Observing System Research and Predictability Experiment; Bougeault et al., 2010), EFAS (The European Flood Awareness System; Thiemig and Salamon, 2017), GloFAS (Global Flood Awareness System; Alfieri et al., 2013), and The Météo-France SIM System (Regimbeau et al., 2007). While there is a growing interest for predictions up to several months into the future, very little is known about the potential predictability of hydrological extremes with annual to decadal lead times, even though they are recognized as an essential tool for near-term planning (IPCC, 2007; Kirtman et al., 2013; Meehl et al., 2009; Vera et al., 2010).

One viable approach to obtain decadal flood forecasts is through the use of statistical models in which the predictand (e.g., the frequency of flood events) is related to relevant predictors, such as precipitation and temperature (e.g., Slater and Villarini, 2018). Data-driven approaches of this kind are recognized to be straightforward to implement, especially given the availability of climate data, and they may yield results as accurate as those obtained by physically-based models (Pagano et al., 2009; Duethmann et al., 2015). One key component in statistical flood forecasting is identifying which physical drivers are responsible for the observed inter-annual variability: there have been growing efforts towards attribution studies of this kind, indicating that precipitation and antecedent moisture conditions are dominant flood drivers (e.g., Mallakpour and Villarini, 2015; Slater and Villarini, 2016, 2017, 2018). Building on the insights gained from these attribution studies, the research questions we aim to answer in this study are:

Is it possible to skillfully predict the frequency of flood events across the central United States? How does the forecast skill vary as the lead time increases from one to ten years? Can the skill be improved through bias correction and statistical downscaling of the precipitation and temperature inputs?

2. Data

Here we use daily streamflow records from 1940 to 2016, from 286 U.S. Geological Survey (USGS) gaging stations within the U.S. Midwest (defined here to include Illinois, Indiana, Iowa, Kansas, Michigan, Minnesota, Missouri, Nebraska, North Dakota, Ohio, South Dakota and Wisconsin). Each time series has the following properties: (i) it has at least 50 consecutive years of measurements; (ii) it contains only complete calendar years (we define the year as complete when it has at least 330 days of measurements); and (iii) its corresponding water basin is not affected by any significant regulation (i.e., no codes “5” or “6” in the USGS records).

To create the time series of flood counts (i.e., the predictand), we use a peak-over-threshold (POT) approach. The flood threshold is defined computationally as the streamflow rate above which only a fixed number of flood days have occurred over the entire period. We set four different thresholds allowing, on average, from 1 to 4 events per year over the 1940-2016 period. To avoid double counting the same event, we only consider one event in a window of ± 5 days plus the logarithm of the drainage area (in square miles) (Lang et al., 1999). Based on these records, we associate each event to a calendar season (winter: December-February; spring: March-May; summer: June-August; fall: September-November), and separate models are developed for each season.

In terms of predictors, we use the PRISM raster dataset (PRISM, 2017), which provides daily time series of climate variables at a resolution of ~4km. Basin-averaged observed daily precipitation and temperature are extracted from PRISM over the period 1940-2016. The predicted precipitation and temperature time series are derived from several GCMs from the fifth generation of the Coupled Model Intercomparison Project (CMIP5) (Taylor et al., 2012, 2007). From this dataset we derive the so called decadal predictions, which are hindcasts and predictions initialized every one to five years (depending on the GCM) from 1960 to 2005 (Meehl et al., 2009). We use the equal-weight ensemble average of multiple GCMs (nine for precipitation and 14 for temperature) and we select lead times (i.e., the lag time between the initialization year and the predicted year) ranging from one to ten years. Because of the different initialization years, the number of members constituting the ensemble can vary from two to nine GCMs for precipitation, and up to 14 GCMs for temperature. We compare results obtained using both the raw GCM data and a statistically downscaled and bias corrected (SD-BC) version of it: among the several techniques to carry out bias correction and statistical downscaling of climate predictions (Choudhury et al., 2017; Li et al., 2010; Panofsky and Brier, 1968; Wood et al., 2004), in this work we used the approach developed by Salvi et al (2017a,b) because it provides high-resolution and bias-corrected predictions of temperature and precipitation for the continental United States.

Salvi et al. (2017a,b) obtained high resolution decadal predictions of monthly temperature and precipitation over the continental United States using two approaches: 1) bias correction; and 2) transfer function based statistical downscaling. For the bias correction of raw decadal temperature predictions produced by GCMs, they applied the Quantile Mapping Technique proposed by Li et al. (2010). The methodology relies on: a) replacing the outputs by the GCMs

over the calibration period with the observed data, so that they have the same cumulative distribution function (CDF) value; and b) computing and retaining the shift between the CDFs of the GCM predictions over the calibration and testing periods. This approach is applied to each grid cell after the GCM predictions are downscaled to the resolution of the observed data using a bilinear interpolation technique; furthermore, it is applied separately for each month and lead time.

The second approach is the transfer function approach, which involves establishing statistical relationships between coarse resolution climate variables (predictors) in the form of reanalysis data and the climate variable of interest (temperature/precipitation). The statistical methodology used by Salvi et al. (2017a) to downscale the precipitation variable involves: a) identification of zone of predictors, which represents the region where predictors reveal the maximum influence over the predictand (i.e., precipitation); b) calibration and validation of the statistical downscaling model with respect to reanalysis data, where the predictors in the zone of influence undergo dimensionality reduction (i.e., principal component analysis) and the relationship between the precipitation variable and predictors is established and validated; and c) application of the established relationship to obtain high resolution predictions. The transfer functions applied in Salvi et al. (2017a, b) are linear regression and kernel regression. In this work we use kernel regression, and this methodology is applied separately to each month and lead time.

For both the observed and predicted climate variables we compute the basin-averaged seasonally-aggregated values to obtain the respective seasonal time series. These data (whether observed or predicted) represent the inputs for the statistical models we use to obtain decadal predictions of the frequency of flood events.

3. Methodology

Given the discrete nature of the predictand (e.g., number of flood events in a given season), Poisson regression represents an appropriate statistical framework to relate the response variable to different predictors. We consider precipitation, wetness conditions (i.e., the precipitation accumulated during the season prior to the one to forecast) and temperature as covariates. The different models are summarized in **Error! Reference source not found.**. The *Mixed* model uses two different temperature indices depending on the chosen season: during summer and fall the *Mixed* model includes the average summer temperature ($x_{T_{Summer}}$) as a predictor, to reflect the effects of evaporation in the summer and drying of the soil in fall; during spring this model considers the average March-April temperature ($x_{T_{Mar-Apr}}$) as a simple way of representing snowmelt processes.

We fit these models to the observed flood count time series over the period 1940-2016 and select the best model according to the Bayesian Information Criteria (BIC) for each station, season and flood threshold value. We eliminate a gage station if either of the two following conditions applies (because both of these conditions can generate implausible model fits): (i) if the station's observed flood count time series has less than five years with a number of events different from zero; (ii) if the parameter reflecting precipitation is negative. Once the best model has been selected for each station, season and flood threshold value, we estimate its parameters using the observed records from 1940 up to the initialization year of the predictions or hindcasts, using the retroactive validation approach described in Mason and Baddour (2007). We then use the decadal predictions of the predictors to obtain the forecasts of the frequency of flood events with a lead time ranging from one to ten years. Given that Salvi et al. (2017a, b) used the 1960-

1990 period to calibrate their bias correction and statistical downscaling approach, we quantify the skill in predicting the frequency of flood events for the validation period ranging from 1991 to 2016.

To quantify the skill of our predictions, we compute the correlation coefficient between the observations and the median of the predicted Poisson distribution. Furthermore, we decompose the actual skill score SS into three components (Murphy and Winkler, 1992):

$$SS = \rho_{po}^2 - \left[\rho_{po} - \left(\frac{\sigma_p}{\sigma_o} \right) \right]^2 - \left[\frac{(\mu_p - \mu_o)}{\sigma_o} \right]^2 \quad (1)$$

where ρ_{po} is the correlation between observed and predicted time series and σ_p (σ_o) and μ_p (μ_o) are the standard deviation and mean of the predicted (observed) time series, respectively. The first term on the right side of the equation is the coefficient of determination, which represents the potential skill that could be obtained in the absence of biases; the second term is the slope reliability (which is an expression of conditional biases) and the third term is the standardized mean error (a measure of unconditional biases). We assess the influence of the biases on the predicted flood count time series by showing both the correlation coefficient and the skill score SS . Note that the period over which the two indexes are evaluated changes according to the lead time: for instance if we consider a lead time of one year, the period ranges from 1991 to 2015; if the lead time is 9 years, the period is 1999-2016.

4. Results and Discussion

As a first step, we examine the skill of the basin-averaged decadal predictions of precipitation and temperature in reproducing the observed seasonal climate values. Figure 1 shows the skill of precipitation forecasts for different lead times using the SD-BC GCM ensemble. The skill does not decrease monotonically as the lead times increase (see also Figure

S1), i.e. longer lead times do not always perform worse than the shorter ones, possibly due to the initial effects associated with the models' initialization (e.g., Salvi et al. 2017a, b). Results indicate that there is not a season, or a region/subregion, that is predicted consistently better than the others: the skill (or lack of skill) is fairly consistent both spatially and temporally. Moreover, compared to the raw predictions (Figure S2), the SD-BC precipitation predictions do not always have a higher skill. The limited skill in decadal predictions of precipitation was also shown, for instance, by Mehrotra et al. (2014) for Australia.

The skill of the temperature forecasts is better than that of precipitation (Figure 2). The correlation coefficients are generally higher and much more spatially homogenous, especially during summer and fall. In contrast to the precipitation, downscaling and bias-correcting the temperature ensemble time series leads to an overall increase of the prediction skill (compare Figure S3 and Figure S4); however, shorter lead times are not always associated with higher values of the correlation coefficient. Overall, the skill in predicting precipitation and temperature at long lead times is comparable with what observed for their seasonal predictions (Supplementary Figures S2-S3 in Slater and Villarini (2018)), suggesting that the challenges in predicting these variables are shared between seasonal and decadal lead times. It is also worth mentioning that the results based on the correlation coefficient do not account for conditional and unconditional biases (equation 1); the SD-BC method is designed to mitigate these biases.

Figures 1-2 indicate that there is some skill in forecasting the climate inputs, especially in terms of temperature. However, from a flood prediction point of view, because precipitation is the dominant flood driver (x_p and x_M), we would be more interested in having a better skill in precipitation forecasts: yet our results indicate that the precipitation skill is low, spatially inconsistent, and varies across seasons and lead times. Figure 3 shows the correlation coefficient

between the observed and predicted flood counts time series using a threshold value that gives two peaks per year on average. Not surprisingly, as a consequence of the low prediction skill of precipitation (Figure 1), there is very limited skill in forecasting the frequency of flood events with a one- to ten-year lead time. This statement holds across seasons, areas and forecast horizon (Figure S5).

The low prediction skill is not due to the performance of the statistical model: this can be verified by comparing the statistical model outputs obtained using observations and decadal predictions. Using the statistical models selected for Figure 3, we compute the difference in the value of the correlation coefficient (between observed and estimated flood counts) obtained using observed precipitation/temperature as predictors and the correlation coefficient obtained using the decadal predictions (Figure 4). The average differences range from 0.45 to 0.69, indicating that it is possible to obtain high values of correlations when the statistical models are forced with observations, and that the limited skill is not due to the model's structure but to the limited skill in predicting the covariates.

In contrast with the correlation coefficient results (Figure 3), the skill further decreases when the evaluation metric accounts for conditional and unconditional biases (skill score shown in Figure 5). In winter especially, the skill score presents low values. For every season and every lead time, the main driver of the decrease in prediction skill tends to be the unconditional (Figure S6) rather than the conditional (Figure S7) bias, suggesting that the average number of flood events predicted by the models differs from the observed number of flood events. Figures 3 and 5 indicate that for many stations the correlation coefficient and the skill score could not be calculated, due to two reasons: (i) the prediction skill of flood counts could not be quantified because the observed time series on which the models are calibrated do not have at least five

years of non-zero values; (ii) the predictions of precipitation exhibited low inter-annual variability (Salvi et al. 2017a), leading to the generation of time series with a constant number of flood counts (for which the standard deviation is zero).

While these results are based on the SD-BC outputs, we obtain even lower skill when we use the raw ensemble. Figure S8 (Figure S9) shows the stacked bar plot of the percentage of the number of stations for which the correlation coefficient (skill score) of the SD-BC ensemble is higher (red part of the bar) and lower (blue part of the bar) than the raw ensemble, for every season, lead time and every threshold value (the percentage is calculated among all the stations for which the two indexes could be assessed). While there is no clear difference in terms of correlation coefficient between the raw and the SD-BC results (Figure S8), the downscaling process and the correction of biases lead to an overall increase of the skill score for every discharge threshold values and lead times, particularly during spring (Figure S9). In terms of the sensitivity of the results to the selected threshold, the correlation coefficient does not seem to change when we consider more frequent events (Figures S5, S10-S12); the skill score, on the contrary, slightly increases when the threshold value decreases (Figures S13-S16), suggesting that focusing on more frequent events reduces the conditional and unconditional biases in predicting flood counts.

5. Conclusions

This study represents the first attempt to produce forecasts of the seasonal frequency of flood events over annual to decadal timescales using a statistical dynamical approach. Using observed and predicted time series of precipitation and temperature at 286 stream gaging stations across the U.S. Midwest, we have examined the skill in predicting the seasonal frequency of

flood events with lead times ranging from one to ten years. The main findings of this work can be summarized as follows:

- The decadal predictions of basin-averaged precipitation exhibit limited skill, regardless of the season, lead time or geographic area. Bias correction and statistical downscaling slightly improve the results, especially for temperature. The decadal predictions of temperature exhibit higher correlation coefficients with respect to the observations, with a more spatially homogeneous skill.
- The limited skill in the input (climate forecasts) limits the skill in the prediction of the frequency of flood events. This becomes even more apparent once we account for conditional and unconditional biases in the relationship between observed and predicted counts.
- The skill in predicting the frequency of flood events is not sensitive to the frequency or rarity of flood events.
- Given the significant role that precipitation plays in flooding and the limited skill we currently have in forecasting rainfall at annual to decadal time scales, it is not surprising that the decadal predictions of flooding are not skillful. Moving forward, our improved capability in predicting the frequency of flood events will stem from our improved predictions of precipitation. Another path forward is to focus on quantities that are related to temperature (e.g., heatwaves), which were shown to be more skillfully predicted at yearly and longer time scales.

Although the skill of streamflow forecasts is currently limited over annual to decadal timescales due to the lack of skill of GCMs, this study provides a benchmark that can be used to

evaluate the progress in predicting precipitation, temperature and the frequency of flood events with new generations of GCMs.

Acknowledgements: This study was supported in part by the Broad Agency Announcement (BAA) Program and the Engineer Research and Development Center (ERDC)–Cold Regions Research and Engineering Laboratory (CRREL) under Contract No. W913E5-16-C-0002, and by the National Science Foundation under CAREER Grant AGS-1349827. The comments and suggestions by two anonymous reviewers are gratefully acknowledged.

References

- Alfieri, L., Burek, P., Dutra, E., Krzeminski, B., Muraro, D., Thielen, J., Pappenberger, F., 2013. GloFAS-global ensemble streamflow forecasting and flood early warning. *Hydrol. Earth Syst. Sci.* 17, 1161–1175. <https://doi.org/10.5194/hess-17-1161-2013>
- Ashley, S.T., Ashley, W.S., 2008. Flood fatalities in the United States. *J. Appl. Meteorol. Climatol.* 47, 805–818. <https://doi.org/10.1175/2007JAMC1611.1>
- Bougeault, P., Toth, Z., Bishop, C., Brown, B., Burridge, D., De Chen, H., Ebert, B., Fuentes, M., Hamill, T.M., Mylne, K., Nicolau, J., Paccagnella, T., Park, Y.Y., Parsons, D., Raoult, B., Schuster, D., Dias, P.S., Swinbank, R., Takeuchi, Y., Tennant, W., Wilson, L., Worley, S., 2010. The thorpe interactive grand global ensemble. *Bull. Am. Meteorol. Soc.* 91, 1059–1072. <https://doi.org/10.1175/2010BAMS2853.1>
- Carsell, K.M., Pingel, N.D., Ford, D.T., 2004. Quantifying the Benefit of a Flood Warning System. *Nat. Hazards Rev.* 5, 131–140. [https://doi.org/10.1061/\(ASCE\)1527-6988\(2004\)5:3\(131\)](https://doi.org/10.1061/(ASCE)1527-6988(2004)5:3(131))
- Choudhury, D., Sen Gupta, A., Sharma, A., Mehrotra, R., Sivakumar, B., 2017. An Assessment of Drift Correction Alternatives for CMIP5 Decadal Predictions. *J. Geophys. Res. Atmos.* 122, 10,282–10,296. <https://doi.org/10.1002/2017JD026900>
- Cloke, H.L., Pappenberger, F., 2009. Ensemble flood forecasting: A review. *J. Hydrol.* 375, 613–626. <https://doi.org/10.1016/j.jhydrol.2009.06.005>
- Downton, M.W., Pielke, R.A., 2005. How accurate are disaster loss data? The case of U.S. flood damage. *Nat. Hazards* 35, 211–228. <https://doi.org/10.1007/s11069-004-4808-4>
- Duethmann, D., Bolch, T., Farinotti, D., Kriegel, D., Vorogushyn, S., Merz, B., Pieczonka, T., Jiang, T., Su, B., Güntner, A., 2015. Attribution of streamflow trends in snow- and glacier melt dominated catchments of the Tarim River, Central Asia. *Water Resour. Res.* 51, 4727–4750. <https://doi.org/10.1002/2014WR016716>.Received
- Garen, B.D.C., 1993. Improved techniques in regression-based streamflow volume forecasting 118, 654–670.
- He, Y., Wetterhall, F., Bao, H., Cloke, H., Li, Z., Pappenberger, F., Hu, Y., Manful, D., Huang, Y., 2010. Ensemble forecasting using TIGGE for the July-September 2008 floods in the Upper Huai catchment: A case study. *Atmos. Sci. Lett.* 11, 132–138. <https://doi.org/10.1002/asl.270>
- IPCC, 2007. Climate Change 2007 Synthesis Report, Intergovernmental Panel on Climate Change [Core Writing Team IPCC. <https://doi.org/10.1256/004316502320517344>
- Kirtman, B., Power, S.B., Adedoyin, J.A., G.J., B., Bojariu, R., Camilloni, I., Doblas-Reyes, F.J., Fiore, A.M., Kimoto, M., Meehl, G.A., Prather, M., Sarr, A., Schar, C., Sutton, R., van Oldenborgh, G.J., Vecchi, G., Wang, H.J., 2013. Near-term Climate Change: Projections and Predictability. *Clim. Chang.* 2013 Phys. Sci. Basis. Contrib. Work. Gr. I to Fifth Assess.

Rep. Intergov. Panel Clim. Chang.

- Kunkel, K.E., Pielke, R.A., Changnon, S.A., 1999. Temporal Fluctuations in Weather and Climate Extremes That Cause Economic and Human Health Impacts: A Review 1077–1098. [https://doi.org/10.1175/1520-0477\(1999\)080<1077:TFIWAC>2.0.CO;2](https://doi.org/10.1175/1520-0477(1999)080<1077:TFIWAC>2.0.CO;2)
- Lang, M., Ouarda, T., Bobée, G., 1999. Towards operational guidelines for over-threshold modeling. *J. Hydrol.* 225, 103–117.
- Li, H., Sheffield, J., Wood, E.F., 2010. Bias correction of monthly precipitation and temperature fields from Intergovernmental Panel on Climate Change AR4 models using equidistant quantile matching. *J. Geophys. Res. Atmos.* 115. <https://doi.org/10.1029/2009JD012882>
- Mallakpour, I., Villarini, G., 2015. The changing nature of flooding across the central United States. *Nat. Clim. Chang.* <https://doi.org/10.1038/nclimate2516>
- Mason, S.J., Baddour, O., 2007. Seasonal Climate: Forecasting and Managing Risk. <https://doi.org/10.1007/978-1-4020-6992-5>
- Meehl, G.A., Goddard, L., Murphy, J., Stouffer, R.J., Boer, G.J., Danabasoglu, G., Dixon, K., Giorgetta, M.A., Greene, A.M., Hawkins, E., Hegerl, G., Karoly, D., Keenlyside, N., Kimoto, M., Kirtman, B., Navarra, A., Pulwarty, R., Smith, D., Stammer, D., Stockdale, T., 2009. Can It Be Skillful? *Bull. Am. Meteorol. Soc.* 90, 1467–1486. <https://doi.org/10.1175/2009BAMS2778.I>
- Mehrotra, R., Sharma, A., Bari, M., Tuteja, N., Amirthanathan, G., 2014. An assessment of CMIP5 multi-model decadal hindcasts over Australia from a hydrological viewpoint. *J. Hydrol.* 519, 2932–2951. <https://doi.org/10.1016/j.jhydrol.2014.07.053>
- Mendoza, P.A., Wood, A.W., Clark, E., Rothwell, E., Clark, M.P., Nijssen, B., Brekke, L.D., Arnold, J.R., 2017. An intercomparison of approaches for improving operational seasonal streamflow forecasts. *Hydrol. Earth Syst. Sci.* 21, 3915–3935. <https://doi.org/10.5194/hess-21-3915-2017>
- Murphy, A., Winkler, R., 1992. Diagnostic verification of probability forecasts. *Int. J. Forecast.* 7, 435–455.
- Pagano, T.C., Garen, D.C., Perkins, T.R., Pasteris, P.A., 2009. Daily updating of operational statistical seasonal water supply forecasts for the Western U.S. *J. Am. Water Resour. Assoc.* 45, 767–778. <https://doi.org/10.1111/j.1752-1688.2009.00321.x>
- Panofsky, H.A., Brier, G.W., 1968. Some Applications of Statistics to Meteorology.
- Piechota, T.C., Chiew, F.H.S., Dracup, J.A., 1998. Seasonal streamflow forecasting in eastern Australia and the el-nino-Southern Oscillation 34, 3035–3044.
- Plummer, N., Tuteja, N., Wang, Q.J., Wang, E., Robertson, D., Zhou, S., Schepen, a, Alves, O., Timbal, B., Puri, K., Division, W., 2009. A seasonal water availability prediction service: opportunities and challenges. 18th World IMACS / MODSIM Congr. 80–94.

PRISM, 2017. Prism Climate Group [WWW Document].

Regimbeau, F.R., Habets, F., Martin, E., Noilhan, J., 2007. Ensemble Streamflow forecasts over France. ECMWF Newsl. No. 111 111, 21–27. <https://doi.org/10.21957/sbjdm2xu33>

Robertson, D.E., Wang, Q.J., 2012. A Bayesian Approach to Predictor Selection for Seasonal Streamflow Forecasting. *J. Hydrometeorol.* 13, 155–171. <https://doi.org/10.1175/JHM-D-10-05009.1>

Salvi, K., Villarini, G., Vecchi, G.A., 2017a. High resolution decadal precipitation predictions over the continental United States for impacts assessment. *J. Hydrol.* 553, 559–573. <https://doi.org/10.1016/j.jhydrol.2017.07.043>

Salvi, K., Villarini, G., Vecchi, G.A., Ghosh, S., 2017b. Decadal temperature predictions over the continental United States: Analysis and Enhancement. *Clim. Dyn.* <https://doi.org/10.1007/s00382-017-3532-1>

Slater, L.J., Villarini, G., 2018. Enhancing the Predictability of Seasonal Streamflow With a Statistical-Dynamical Approach. <https://doi.org/10.1029/2018GL077945>

Slater, L.J., Villarini, G., 2017. Evaluating the drivers of seasonal streamflow in the U.S. Midwest. *Water (Switzerland)*. <https://doi.org/10.3390/w9090695>

Slater, L.J., Villarini, G., 2016. Recent trends in U.S. flood risk. *Geophys. Res. Lett.* 43, 12,428–12,436. <https://doi.org/10.1002/2016GL071199>

Taylor, K.E., Stouffer, R.J., Meehl, G.A., 2012. An overview of CMIP5 and the experiment design. *Bull. Am. Meteorol. Soc.* 93, 485–498. <https://doi.org/10.1175/BAMS-D-11-00094.1>

Taylor, K.E., Stouffer, R.J., Meehl, G. a, 2007. A Summary of the CMIP5 Experiment Design. *World 4*, 1–33. <https://doi.org/10.1175/BAMS-D-11-00094.1>

Thiemig, V., Salamon, P., 2017. European Flood Awareness System. *Jt. Res. Cent.*

van der Knijff, J.M., Younis, J., De Roo, A. P. J., 2010. LISFLOOD: a GIS--based distributed model for river basin scale water balance and flood simulation. *Int. J. Geogr. Inf. Sci.* <https://doi.org/10.1080/13658810802549154>

Vera, C., Barange, M., Dube, O.P., Goddard, L., Griggs, D., Kobysheva, N., Odada, E., Parey, S., Polovina, J., Poveda, G., Seguin, B., Trenberth, K., 2010. Needs Assessment for Climate Information on Decadal Timescales and Longer. *Procedia Environ. Sci.* 1, 275–286. <https://doi.org/10.1016/J.PROENV.2010.09.017>

Wood, A.W., Lettenmaier, D.P., 2006. A test bed for seasonal hydrologic forecasting approaches in the western united states. *Am. Meteorol. Soc.* 1699–1712. <https://doi.org/10.1175/BAMS-87-I2-I699>

Wood, A.W., Leung, L.R., Sridhar, V., Lettenmaier, D.P., 2004. Hydrologic implications of dynamical and statistical approaches to downscaling climate model outputs. *Clim. Change*

Model Name	Model Hypothesis	Model Predictors
<i>P</i>	$\log \lambda_1 = \alpha_1 + \beta_1 x_P$	x_P - Precipitation
<i>P.T</i>	$\log \lambda_2 = \alpha_2 + \beta_2 x_P + \gamma_2 x_T$	x_P - Precipitation x_T - Temperature
<i>P.M</i>	$\log \lambda_3 = \alpha_3 + \beta_3 x_P + \gamma_3 x_M$	x_P - Precipitation x_M - Wetness Conditions
<i>Mixed</i>	$\log \lambda_4 = \alpha_4 + \beta_4 x_P + \gamma_4 x_M + \delta_4 x_{T_{Mar-Apr}}$ $\log \lambda_4 = \alpha_4 + \beta_4 x_P + \gamma_4 x_M + \delta_4 x_{T_{Summer}}$	x_P - Precipitation x_M - Wetness Conditions x_T - Temperature

Table 1 – List of the statistical models used to relate the seasonal occurrence of flood events to the three drivers (precipitation, wetness conditions and temperature).

156x65mm (300 x 300 DPI)

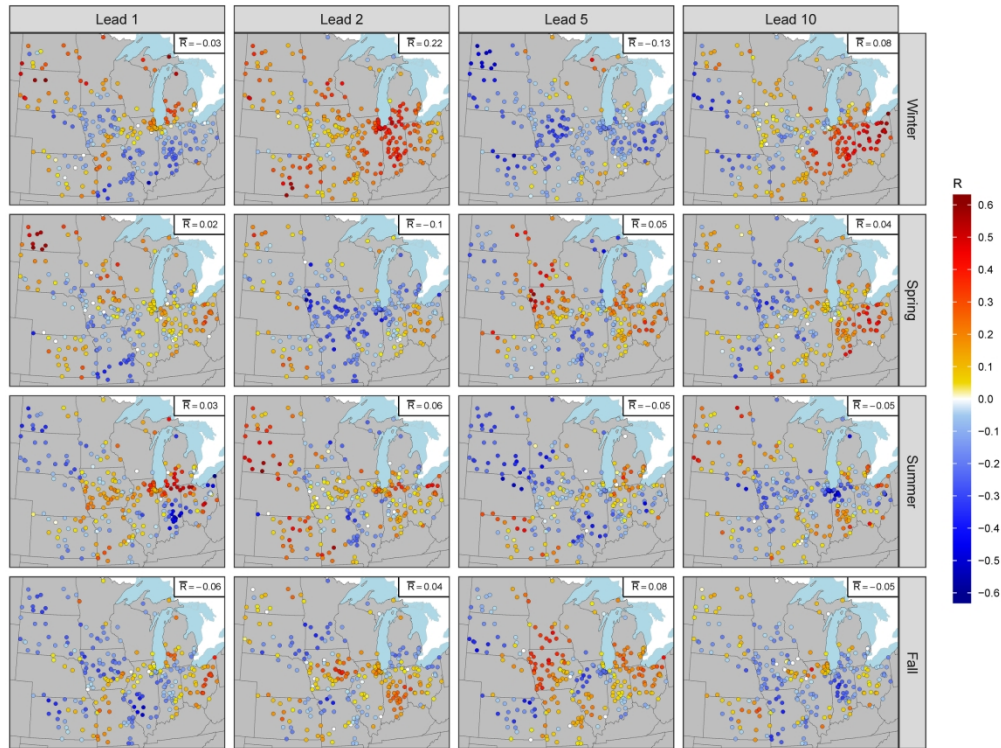


Figure 1 - Map showing the correlation coefficient between observed and predicted precipitation time series. The predicted values are obtained from the SD-BC GCMs ensemble performed by Salvi et al. (2017a). Each row corresponds to a season and each column to a lead time. The same figure for all the lead times is presented in the supplemental material (Figure S1). The average value of the correlation coefficient among all stations is shown in the top right corner of each panel.

223x168mm (300 x 300 DPI)

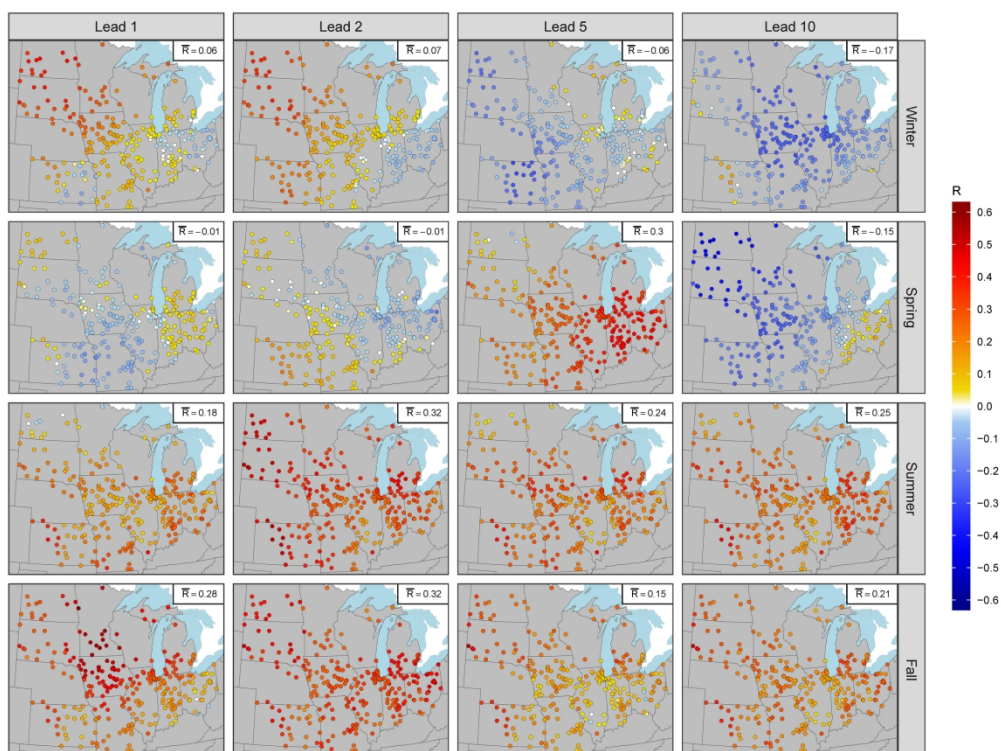


Figure 2 - Map showing the correlation coefficient between observed and predicted temperature time series. The predicted values are obtained from the SD-BC GCMs ensemble performed by Salvi et al. (2017b). Each row corresponds to a season and each column to a lead time. The same figure for all the lead times is presented in the supplemental material (Figure S3). The average value of the correlation coefficient among all stations is shown in the top right corner of each panel.

223x168mm (300 x 300 DPI)

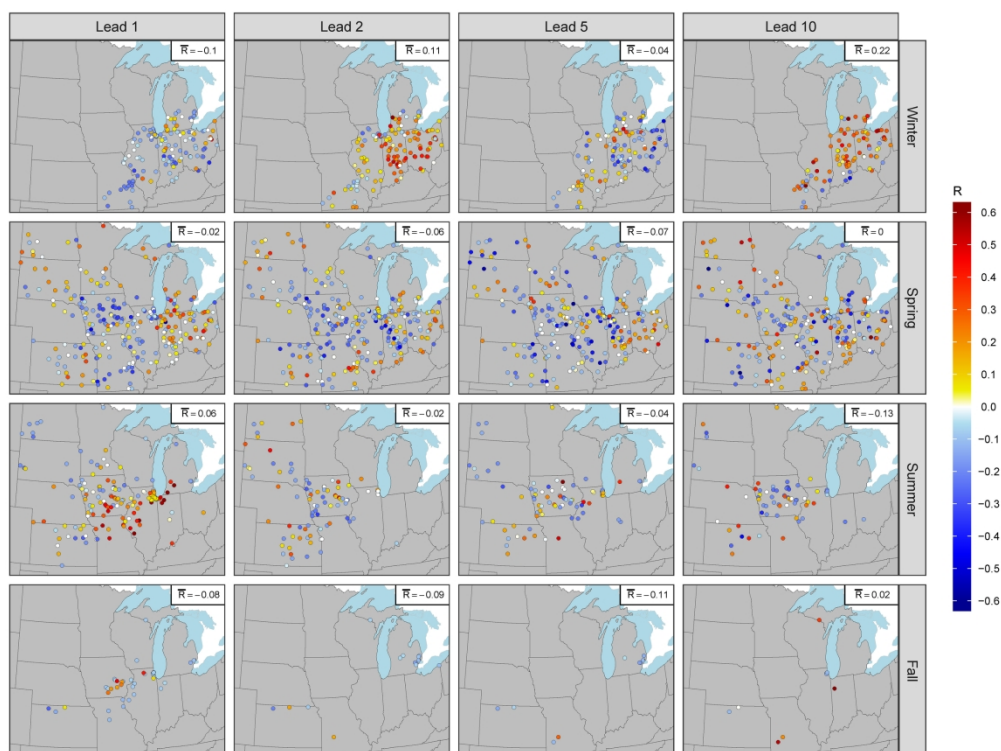


Figure 3 – Map showing the correlation coefficient between observed and predicted flood counts time series using the SD-BC GCMs ensemble. Each panel correspond to a lead time (on the columns) and to a season (on the rows). The threshold value used to extract the observed flood counts time series is defined to give two peaks per year on average. The average value of the correlation coefficient among all stations is shown in the top right corner of each panel. Only stations that have at least five years of non-zero flood counts and that do not have a constant value across the period of record are shown.

223x168mm (300 x 300 DPI)

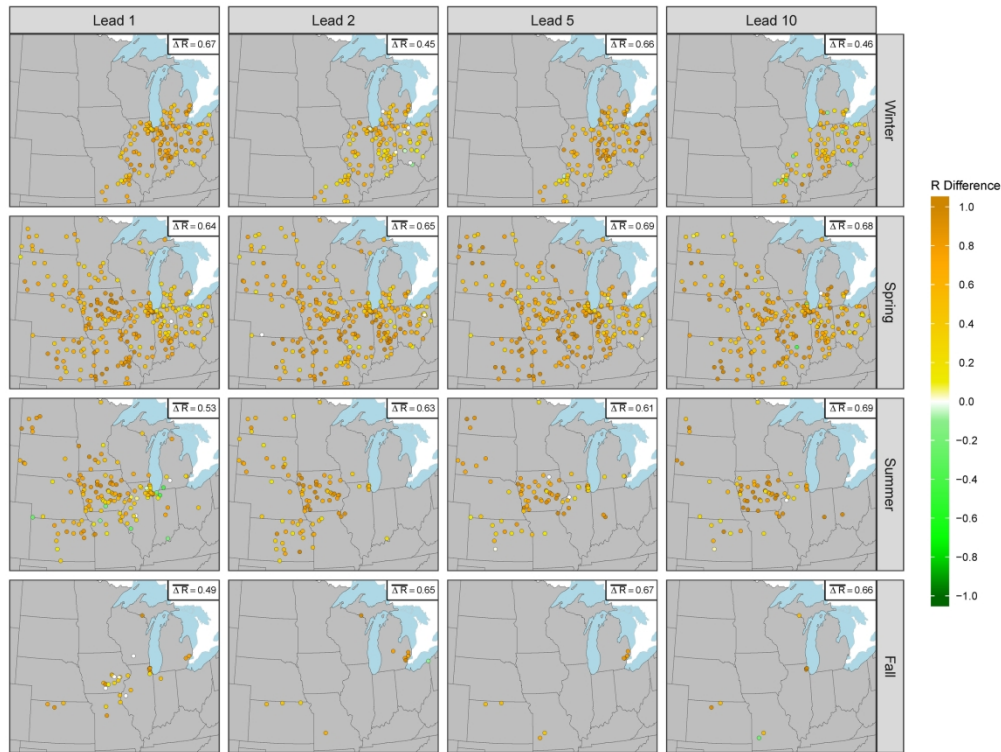


Figure 4 – Map showing the difference between the correlation coefficient between observed and estimated flood counts calculated using observed precipitation/temperature and the correlation coefficient between observed and estimated flood counts based on the SD-BC GCM predictions of precipitation/temperature. Each panel correspond to a lead time (on the columns) and to a season (on the rows). The threshold value used to extract the observed flood counts time series is defined to give two peaks per year on average. The average value of the difference between the two correlation coefficient among all stations is shown in the top right corner of each panel. Only stations that have at least five years of non-zero flood counts and that do not have a constant value across the period of record are shown.

223x168mm (300 x 300 DPI)

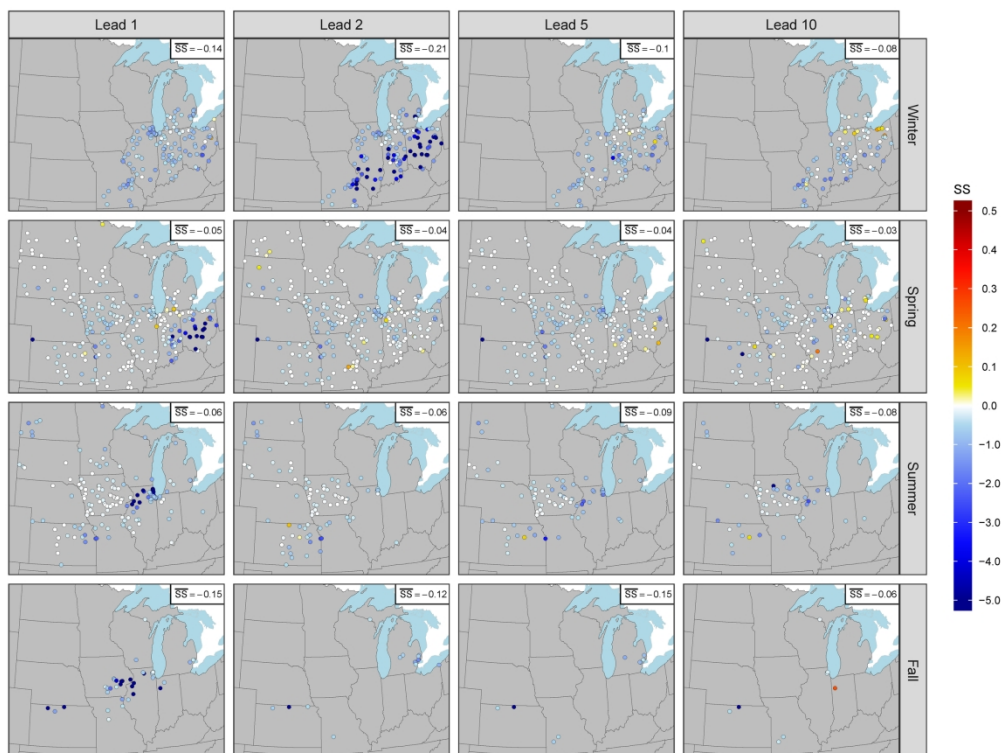


Figure 5 – Same as Figure 3 but using the skill score instead of the correlation coefficient to evaluate the prediction skill. Only stations that have at least five years of non-zero flood counts and that do not have a constant value across the period of record are shown.

223x168mm (300 x 300 DPI)