Accounting for spatial displacement errors in HRRRE QPF to create short-term ensemble streamflow forecasts

by

Kyle Hugeback

A thesis submitted to the graduate faculty

in partial fulfillment of the requirements for the degree of

MASTER OF SCIENCE

Major: Meteorology

Program of Study Committee:
Kristie J. Franz, Co-major Professor
William A. Gallus Jr., Co-major Professor
Chris Rehmann

The student author, whose presentation of the scholarship herein was approved by the program of study committee, is solely responsible for the content of this thesis. The Graduate College will ensure this thesis is globally accessible and will not permit alterations after a degree is conferred.

Iowa State University

Ames, Iowa

2021

TABLE OF CONTENTS

Page

LIST OF FIGURES

Page

LIST OF TABLES

Page

GLOSSARY OF TERMS

Base HRRRE: Streamflow ensemble using the QPFs (see QPF) of the 9 original members of the HRRRE as forcing.

CI: (Convective Initiation) The time and location when a precipitation system first produces measurable rainfall.

CI Displacement: The location of the centroid of the spatially displaced HRRRE QPFs at CI (see CI). It is found by comparing the centroid, or center of mass of the observed precipitation areas and the centroid of the HRRRE member QPF precipitation areas in the Upper Mississippi Valley region.

CI Displacement Weighting: A weighting scheme based on the inverse of the distance between the CI displacement location (see CI), and the location of the ensemble QPF members within our model grid. Referred to in tables and figures as "CI Weighted".

Closest A: An ensemble formed by selecting the nine members from the full ensemble (see Full Ensemble) by first grouping shifted QPF (see QPF) members with their parent HRRRE model member, then choosing the member with the smallest distances CI displacement (see CI displacement). After the members were selected, they were all given equal weighting.

Closest B: As in the Closest A however, the shifted ensemble members are not grouped with their parent HRRRE model member.

Corrected CI Displacement Weighting: A weighting scheme based on the inverse of the distance between the observed CI location (see CI), after it has been adjusted by the Kiel et al. (2021) correction, and the location of the ensemble QPF members within our model grid.

Full Ensemble: Streamflow ensemble that uses forcing from 54 shifted QPF (see QPF) members, and the 9 original HRRRE QPF members (see HRRRE). More specifically, there are 6 shifted QPF members generated for every single HRRRE member.

HRRRE: (High-Resolution Rapid Refresh Ensemble) Convection-allowing ensemble with nine members (Dowell et al. 2018; Dowell 2020). Source of QPF (see QPF) for informed shifting and streamflow ensemble comparison.

NWM: (National Water Model) Operationally used distributed hydrologic model (Gochis et al. 2020). The NWM version 2.0 configuration is used when running the WRF-Hydro hydrologic model for this study. May be used to refer to the model domain of study as the "NWM domain" or "NWM cutout."

POE: (Probability of Exceedance) Probability that flooding will occur. Found by summing the weights of the ensemble members that predicted a magnitude of peak discharge greater than minor flood stage. When used for deterministic forecast verification, ten thresholds

of POE are used to stratify the verification measures based on ensemble predictive certainty.

snRNG: (Skewed-Normal Random Number Generator) Modified random number generator used to select random locations within the model grid to shift QPF (see QPF) based on the climatological spatial displacement dataset for the HRRRE (see HRRRE) found by Kiel et al. (2021).

ACKNOWLEDGMENTS

I would like to thank my co-major professors, Dr Kristie Franz and Dr. William Gallus, as well as my committee member Dr. Chris Rehmann, for their guidance in this process. I would also like to thank my fellow graduate students, Elizabeth Tirone, Justin Covert, Ezio Mauri, Jon Thielen, and many others for their unending help, and moral support. Lastly, I would like to thank my family for fostering my love of meteorology and hydrology from a young age and encouraging me to follow my passion into higher education.

ABSTRACT

Most of the heavy precipitation that falls in the Upper Mississippi Valley occurs during the period of May to September. Most of that warm season precipitation falls as a result of mesoscale convective systems. Because convective rainfall is usually localized and intense, it is critical to account for any model error resulting in spatial displacements. This is especially important when producing hydrologic forecasts as those displacements cause rainfall to be moved across watershed boundaries. Previous research attempted to shift quantitative precipitation forecasts (QPFs) as input to a hydrologic model to expand streamflow ensembles to better account for spatial displacements of QPF. Climatological spatial displacement errors have been computed for multiple convection-allowing ensembles, including the High-Resolution Rapid Refresh Ensemble (HRRRE), paving the way for the present work to apply those findings to hydrologic ensemble forecasts. A skewed-normal random number generator was used to select locations within the range of the climatological spatial displacement errors identified in the previous work to which the individual members of the HRRRE precipitation forecasts were shifted. Every HRRRE member was shifted 6 times, creating a full ensemble of 63 model members. The full ensemble of QPF and the original HRRRE members were used as forcing for the WRF-Hydro v5.1.1 running in a National Water Model 2.0 configuration. Evaluation of the magnitude of peak discharge, timing of peak discharge, and differing ensemble performance among small and large watersheds were carried out using 50 stream gauges, over 29 storm events. Verification of the magnitude of peak discharge was done by using containment ratio, ranked probability score, and reliability as metrics of probabilistic forecast skill. Deterministic evaluation was conducted using critical success index, equitable threat score, and Heidke skill score, at ten probability of exceedance thresholds.

In addition to using equal weights among model members, two weighting schemes were added to modify the influence of ensemble members based on the observed spatial displacement of HRRRE QPF at convective initiation (CI). Previous research had calculated centroids, or center of mass of the precipitation systems, for the HRRRE QPFs and observed rainfall for 30 storm events in the Upper Mississippi region. The centroids of the modelled rainfall systems were compared to the centroids of observed precipitation areas where there was at least 1mm of rainfall at CI. Within the framework of this study, the centroids of HRRRE QPFs at CI, taken from the prior research, were compared to the centroids of the shifts produced by the random number generator. Weights would then be assigned as the inverse of the distance between those centroids. One of the CI displacement weighting schemes used a climatology-based correction to adjust the location of the displaced HRRRE QPF centroid. That correction adjusted the displacement of HRRRE QPF observed at CI to match the displacement of the centroid of accumulated precipitation more closely for the full forecast event. The second of the two weighting schemes disregarded the correction and did not adjust the magnitude of the spatial displacement of the HRRRE QPF when finding the inverse distance weights.

The ensemble weighting methods allowed for the creation of two ensembles with 9 members and were designed to test an ensemble of similar size to the HRRRE, but with added information about the predicted location of QPF at CI. One of the two new ensembles was made by selecting the members that had the smallest distances between the centroid of the shifted QPF and the centroid of the CI displacements. The other new ensemble was similar; however, one member was chosen out of each parent HRRRE member group. All members in the new 9 member ensembles were given equal weight, and effectively zero weight was assigned to the members that were not selected for the smaller ensemble.

CHAPTER 1. GENERAL INTRODUCTION

1.1 General Introduction

Fatalities due to flooding are the second most prominent source of weather-related fatalities. The 30-year average of flooding deaths per year in the United States is 88, while the 10-year average is even higher at 99 fatalities per year (National Weather Service 2019). Due to this high risk, there is significant effort to improve the forecasting of streamflow, and the flood models themselves. Improving hydrologic forecasting during the warm season is the main focus of this study, because most heavy precipitation in the Upper Mississippi Valley comes during that period (Fritsch et al. 1986; Gallus 2012; Haberlie and Ashley 2019). A majority of warm season precipitation occurs during 10 of the wettest days of the year, distributed throughout the season (Pryor et al. 2009), with 50% of the annual precipitation coming as a result of mesoscale convective systems (MCSs; Haberlie and Ashley 2019).

Estimations of convective precipitation are supplied as forcing to hydrologic models by either quantitative precipitation estimates (QPEs) or quantitative precipitation forecasts (QPFs). QPEs are made by processing observational data from rain gauges, radar data, satellite returns, or a combination of any of these components. They can be produced up to near real time for ongoing rainfall events. To be able to provide more lead time for risk mitigation and emergency management, there is large push for the use of QPFs in the creation of hydrologic forecasts (Vasiloff et al. 2007), including their operational use at River Forecast Centers across the United States (Adams 2016). QPFs have been improving over a few decades for many reasons; however, this is mainly due to the increases in computing capabilities and the use of finer grid spacing, and eventually convection-allowing models (CAMs; Seo et al. 2018). CAMs have shown particular improvements over their convectively parameterized counterparts in their

depiction of the diurnal precipitation patterns (Clark et al. 2009; Berenguer et al. 2012), as well as in the representation of convective mode (Kain et al. 2006).

Even though QPF skill has increased, problems still arise when QPFs, which are outputs of atmospheric models, are used to drive a hydrologic model. Errors from any point in the modelling process can propagate throughout the rest of the system and into the streamflow forecast (Brown and Heuvelink 2005; Collier 2007). Because a single deterministic model run is more prone to these sources of error, ensembles made up of multiple model members are increasingly used to account for these errors (Du et al. 1997; Ebert 2001). Ensembles are often created by using perturbed initial and lateral boundary conditions within the same model core, but they can also make use of different physical parameterization schemes, or even multiple model cores. Even more recently, the development of convection-allowing ensembles (CAEs) has allowed researchers to produce probabilistic forecasts with the continued benefits of CAMs.

Spatial displacement errors continue to be one of the major shortcomings in modelling of convection and MCSs, and these are problematic for streamflow modelling (Clark et al. 2009; Gallus 2010, 2012; Yan and Gallus 2016; Carlberg et al. 2020; Goenner et al. 2020; Kiel et al. 2021; Viterbo et al. 2020). Attempts have been made to account for spatial displacement errors in QPFs used in streamflow ensembles through systematic shifting of QPF forcing, with mixed results (Carlberg et al. 2020). Carlberg et al. (2020) used systematic shifts of 0.5 degree and 1 degree latitude and longitude shifts of QPF. Those shifts were made in the four cardinal directions and four intermediate cardinal directions. Although this shifting method increased the ensemble's ability to detect flooding potential, the systematic shifts contributed to a large amount of zero-QPF ensemble members. To further increase the usability of CAMs and CAEs, Kiel et al. (2021) sought to quantify the climatological spatial displacement error for two different CAEs.

Using these climatological spatial displacement errors from Kiel et al. (2021), and the functionality of spatial shifting of QPF as input for hydrologic modelling seen in Carlberg et al. (2020), the present work implements an informed QPF shifting technique to construct streamflow ensemble forecasts for a region of the Upper Mississippi Valley using the WRF-Hydro hydrologic model in the recently updated National Water Model 2.0 configuration (Gochis et al. 2020).

1.2 Research Question and Thesis Format

This research aimed to test if implementing randomized spatial shifting of QPF forcing, informed by the climatology of the accumulated spatial displacements found in Kiel et al. (2021), could improve ensemble streamflow forecasts for flood events. This was done by creating and comparing QPF-driven ensemble streamflow forecasts for 50 stream gauges, and 29 storm events over the 2018 warm season in the Upper Mississippi Valley. Evaluations of the streamflow ensembles were broken up into three separate categories by adding schemes for weighting ensemble members based on observed spatial displacements at the hour of convective initiation. Multiple probabilistic verification methods were used, as well as a handful of deterministic measures evaluated at 10 probability of exceedance thresholds.

This thesis follows a journal paper format. Chapter 1 explores a general introduction, research questions, and the format of this work. Chapter 2 contains a literature review of previous research done on extreme precipitation events, as well as ensemble hydrometeorological and flood forecasting. Chapter 3 is a journal article that will be submitted to the Journal of Hydrometeorology, that lays out this research in full. Chapter 4 will cover general conclusions and future work.

CHAPTER 2. LITERATURE REVIEW

Precipitation events over the upper Midwest are heaviest during the warm season (April-September) (Fritsch et al. 1986; Gallus 2012; Haberlie and Ashley 2019). Among those warm season precipitation events, mesoscale convective systems (MCSs) are considered to be some of the most intense. The American Meteorological Society Glossary of Meteorology defines an MCS to be "A cloud system that occurs in connection with an ensemble of thunderstorms and produces a contiguous precipitation area on the order of 100 km or more in horizontal scale in at least one direction." These events are also characterized to be partially self-sustaining with a large areal coverage (American Meteorological Society 2012). MCSs pose a multitude of threats to human life and society at large, with severe wind, hail, and tornadoes all being common. MCSs and the deep, widespread convection associated with them can result in air traffic to be rerouted, and intense localized rainfall leads to flash flooding. In fact, over 50% of annual rainfall in the Midwest and Eastern U.S. can be attributed to MCSs (Haberlie and Ashley 2019). Typically, 30% of that annual rainfall falls within only 10 days, representing the 10 wettest days of the year (Pryor et al. 2009). Climate data suggests that there could be increases in these extreme precipitation events, increasing the hydrological risk posed by these large convective systems (Hejazi and Markus 2009; Andresen et al. 2012).

The hydrological consequences of such intense rainfall events are apparent. Forecasting of these events has been accomplished using quantitative precipitation estimates (QPEs) and/or quantitative precipitation forecasts (QPFs) as input to hydrologic models. QPEs are made by taking observations of precipitation, sometimes from multiple sources, to compile an area-wide estimate of said precipitation. Common sources of QPE inputs are radar-based, rain gauge, and satellite observations. QPEs are available for past events up to near real-time. QPEs have

evolved over time, and today there is a variety of methods and algorithms used to form data into usable inputs for hydrologic models. As part of their work, Seo et al. (2018) used QPE as forcing for the hillslope-model to recreate an eastern Iowa flood event from 2016. They used two forms of QPE, one derived from radar precipitation estimates, while the other was a gauge-corrected Multi-Radar Multi-Sensor (MRMS) product. Both systems performed quite well at simulating the streamflow for this flood event for the Cedar River at two gauging points. In Zhang et al. (2014), they combined radar-based precipitation estimates, rain gauge data, and orographic precipitation climatology. In the Midwest, the orographic climatology method struggled, while good radar coverage allowed that component of QPE to do well. By combining all the of those QPE products, to form a single output, they saw their best skill (Zhang et al. 2014). QPE has shown itself to be useful in hydrologic forecasting, but it is a product only available up to the present time.

Unknown quantities of future precipitation contribute a large amount of uncertainty to streamflow modelling. QPFs can provide estimates of future precipitation to increase lead time for threat mitigation (Vasiloff et al. 2007), and is available in areas where observational networks may be sparse. QPFs generally come as an output of an atmospheric model, with a wide range of modelling sources/agencies. QPFs have been and continue to be used operationally by National Weather Service River Forecast Centers in their streamflow prediction (Adams 2016). Adams (2016) found that deterministic QPF being used for streamflow forecasting in the Midwest showed skill over zero-QPF input. Gallus (2010) used two object-based verification systems to investigate object parameters and how they compared to ensemble skill and spread. In that work, a 5-member ensemble, run at 4 km grid spacing showed good skill in forecasting rain rate, with less than 10% error found for all 6-hour forecasts periods. The smallest errors in rain rate

occurred towards the beginning of the forecast period. The ensemble members struggled to forecast rain area the most, which contributed to errors in total rain volume (Gallus 2010). Moreover, spatial displacement errors for the precipitation regions were quite large. With the standard deviation of those displacement errors being over 50% of their areal scale. In a more recent study, it was found that for a mesoscale region in Maryland, QPF forecasts from the HRRR model performed quite well, despite notable spatial displacement issues (Viterbo et al. 2020). It is important to note that all atmospheric and hydrologic models produce some errors and/or bias, meaning that when those two forecasting methodologies are used in combination, i.e., when QPF is driving streamflow modelling, these errors can aggregate (Brown and Heuvelink 2005; Collier 2007). QPF with longer lead times contributed to greater uncertainty and error in streamflow forecasts (Seo et al. 2018; Lin et al. 2005). Even then, errors accumulated for shorter lead time QPF as the forecast period increased. Some studies pointed to error when QPF was used as forcing beyond 3-6 hours flood forecasts for basins with short response times having significantly higher streamflow errors (Seo et al. 2018; Adams and Dymond 2019).

Understanding how precipitation forecasts are generated within an atmospheric model and the type of model they originate from is critical to diagnosing their error and/or biases. Atmospheric models with large grid spacing struggle to accurately model precipitation and cloud processes due to their dependence on cloud parameterization schemes. The use of cloud parameterization schemes can introduce large errors to precipitation forecasts (Gallus 2012). Only when the model grid spacing drops below 3 km, can this dependence be eliminated; these models are commonly called convection-allowing models (CAMs). Though they still rely on microphysics parameterizations and boundary layer schemes, their forecasting of convective

precipitation is greatly enhanced. CAMs showed greater skill in the diurnal variations in precipitation (Clark et al. 2009; Berenguer et al. 2012), as well as, discerning convective mode (Kain et al. 2006). CAMs have been shown to systematically outperform their convectively-parameterized counterparts, especially in QPFs (Iyer et al. 2016).

A convection-allowing ensemble (CAE) is made up of two or more CAMs. They have been proven to better account for uncertainty in model outputs over deterministic single member models (Du et al. 1997; Ebert 2001). The spread of ensemble members can be formed by a combination of diverse model cores and physics packages, or by using a single model core and providing it perturbed initial conditions and/or lateral boundary conditions. The High-Resolution Ensemble Forecast version 2 (HREF) is an example of an ensemble formed by several deterministic models with different model cores. It is made up of current and time lagged members of the Advanced Research Weather Research and Forecast model (WRF-ARW), the Nonhydrostatic Multiscale Model on the B-grid, "National Severe Storms Laboratory-like"-ARW, and the North American Mesoscale Forecast System. An example of CAE formed using mixed initial conditions is the HRRR Ensemble (HRRRE), which uses perturbed temperature, wind vectors U and V, water vapor mixing ratio fields, as well as 15-minute radar data assimilation to initialize its nine model members (Dowell et al. 2018; Dowell 2020). A direct comparison of these two CAEs found that the HREF had better storm placement than the HRRRE (Roberts et al. 2020). On the other hand, the HRRRE had lower precipitation amounts associated with all probability of exceedance values than the HREF. Those precipitation amounts contributed to the HRRRE precipitation forecasts being closer to the observed precipitation, leading to better streamflow discharge prediction (Goenner et al. 2020).

Even with the improvements afforded to researchers through the use of CAMs and CAEs, spatial displacements of convection continue to be a difficult hurdle to overcome (Clark et al. 2009; Gallus 2010, 2012; Yan and Gallus 2016; Carlberg et al. 2020; Goenner et al. 2020; Kiel et al. 2021; Viterbo et al. 2020). Carlberg et al. (2020) attempted to account for displacement errors using systematic shifting of HRRRE inputs to the Hydrology Laboratory-Research Distributed Hydrologic Model (HL-RDHM). They used a series of 81 half degree and one degree latitude and longitude shifts, in the 8 cardinal directions to expand their ensemble. They concluded that this method for shifting precipitation inputs to the HL-RDHM improved the ability for a forecaster to detect flooding potential. However, due to a large sample of zero runoff shifts produced by the system, probability of exceedance values were low, which did not lend to improved forecast certainty (Carlberg et al. 2020). In conjunction to the previously mentioned study, Kiel et al. (2021) set out to quantify the spatial displacements of the HREF and HRRRE for 30 cases during the 2018 warm season. This was done by setting a forecasted precipitation threshold of 1mm to identify areas that were receiving rainfall. Then a centroid was found for that rain area and compared to the centroid of observed precipitation. Spatial displacements were studied for both the time of convective initiation (CI) and the accumulated 0 to 18-hour displacement. They found that the HRRRE had a smaller spread of climatological displacements at CI than the HREF, while the HRRRE also saw a systematic bias to the west for the 0-18 hour accumulations (Kiel et al. 2021).

CHAPTER 3. ACCOUTING FOR SPATIAL DISPLACEMENT ERRORS IN HRRRE QPF TO CREATE SHORT-TERM ENSEMBLE STREAMFLOW FORECASTS

Kyle K. Hugeback, Kristie J. Franz, William A. Gallus Jr.

Iowa State University Department of Geological and Atmospheric Sciences

Modified from a manuscript to be submitted to the Journal of Hydrometeorology

3.1 Abstract

Errors associated with the location of precipitation in quantitative precipitation forecasts (QPFs) present challenges when they are used for hydrologic prediction, particularly in small basins. Past research has used systematic shifting of QPF for the generation of streamflow ensembles, with increased detection of flood potential, but with low predictive certainty. The present research makes use of a recent climatology of spatial displacement errors in convective-allowing ensembles to perform a more informed shifting of QPF to the individual members of the High-Resolution Rapid Refresh Ensemble. Streamflow predictions for this research were done using the WRF-Hydro version 5.1.1 in a National Water Model 2.0 configuration. Beyond the use of equal weighting for ensemble members, two schemes were added that adjusted member weights based on the spatial displacement of QPF at convective initiation (CI). Three new streamflow ensembles were created: a 63-member ensemble (54 shifted and 9 original members), and two nine member ensembles that were produced using members with the smallest observed spatial displacements at CI. The ensembles using climatologically shifted QPF forcing showed better probabilistic forecasting skill, while having comparable deterministic forecasting skill to the original HRRRE ensemble.

3.2 Introduction

In the upper Midwest, the heaviest precipitation events of the year occur during the warm season (Fritsch et al. 1986; Gallus 2012; Haberlie and Ashley 2019). Warm-season convective

storms, and more specifically, mesoscale convective systems (MCSs), come with many hazards including, severe wind, hail, tornadoes, and intense localized precipitation. MCSs account for 50% of annual rainfall in the Midwest and Eastern U.S. (Haberlie and Ashley 2019). With future climate projections indicating a risk for increased precipitation from these large convective complexes (Hejazi and Markus 2009; Andresen et al. 2012), being able to accurately model these events is a focus of current research. As intense convective precipitation events become more common, being able to accurately model the runoff and streamflow of such events is extremely important.

The socio-economic impacts of flooding and flash flooding cannot be overstated and thus risk mitigation is constantly being explored. Research to improve modelling has investigated the meteorological data that are ingested in hydrologic models. There are two common inputs used as forcing for hydrologic prediction: Quantitative precipitation estimates (QPEs) and quantitative precipitation forecasts (QPFs). QPEs are usually made using satellite observations, radar-based estimation, and/or rain gauge data. QPE has been shown to provide sufficient input to hydrologic model to produce accurate streamflow forecasts (Seo et al. 2018; Zhang et al. 2014). However, QPE is dependent on observed data, thus it is only available for past rainfall events up to near real-time. This lends to QPFs having one major advantage over QPEs, that being an opportunity for increased forecast lead time (Vasiloff et al. 2007). QPFs are generated using atmospheric models for weather prediction. When QPF is used to force hydrologic models, any biases from the weather model will propagate and interact with biases in the hydrologic model (Brown and Heuvelink 2005; Collier 2007). QPFs with longer lead times generally contribute to greater uncertainty and error in streamflow forecasts (Seo et al. 2018; Lin et al. 2005). Even if QPF is provided with little lead time to the start of the rainfall event, errors can accumulate throughout

the hydrologic forecast. Past research has found that forecasts for small watersheds saw significant increases in streamflow error beyond a forecast of 6 hours (Seo et al. 2018; Adams and Dymond 2019).

Operationally, the National Weather Service River Forecast Centers use QPFs to generate some of their forecast products (Adams 2016). In a study that used object-based verification to examine ensemble skill and spread with a comparison to rainfall observations, there was less than a 10% error in forecasting rain rate in a 5-member convection-allowing ensemble, but it struggled with the forecasting of rain area (Gallus 2010). Gallus (2010) also found that spatial displacement errors could be several magnitudes large than the areal scale of the precipitation system. A recent study of a 2018 flood event showed that QPF from the High-Resolution Rapid Refresh (HRRR) model performed very well, with most areal precipitation volumes matching MRMS data (Viterbo et al. 2020). Though the QPF intensities were accurately depicting the risk of flooding in their mesoscale domain of study, spatial displacement errors caused issues with the placement of precipitation at finer scales. Those issues being that the overall average precipitation over a large area was correct, however localized areas of wet biases next to dry biases showed where spatial displacements of convective precipitation had occurred.

Further studies have investigated the issues associated with the forecasting of intensity and location of convective precipitation. Gallus (2012) found that convective parameterization schemes, used in atmospheric models with coarse grid spacing, lead to the failure of models to produce accurate precipitation forecasts during the warm season. Convection-allowing models (CAMs) partially solve this issue by operating on fine grid scales. CAMs have been shown to improve skill in the depiction of convective mode (Kain et al. 2006), while also having a better representation of variations in diurnal precipitation (Clark et al. 2009; Berenguer et al. 2012).

CAMs, though still constrained by microphysics parameterizations, manage to outperform their convectively parametrized counterparts in their generation of QPFs (Iyer et al. 2016).

When multiple CAM members are combined or created, they can be formed into a convection-allowing ensemble (CAE). Past studies have determined that ensembles and the probabilistic forecasts that they create provide a better assessment of flood risk than single member deterministic forecasts (Du et al. 1997; Ebert 2001). Ensembles of opportunity formed by diverse model cores and physics schemes, and traditional ensembles driven by perturbed initial conditions (ICs) using one model core, have both been utilized for hydrologic forecasting. Goenner et al. (2020) found that the High-Resolution Ensemble Forecast version 2 (HREF), with its 4 separate cores and time lagged members, had better storm placement than the High-Resolution Rapid Refresh ensemble (HRRRE), a more traditional ensemble. However, the probability of exceedance values from the HRRR QPF allowed for the generation of forecasts of lower discharge values than the HREF, which were closer to observations (Goenner et al. 2020).

Throughout all of these applications of QPF, spatial displacement errors have been a continued issue (Clark et al. 2009; Gallus 2010, 2012; Yan and Gallus 2016; Carlberg et al. 2020; Goenner et al. 2020; Kiel et al. 2021; Viterbo et al. 2020). Spatial displacements in QPF can be extremely detrimental to hydrologic forecasts. That can be amplified depending on the size and shape of the watershed(s) being studied. As an example, consider a hypothetical watershed that is long in the north-south direction and narrow in the east-west direction. If QPF is displaced in the east-west direction, the watershed may receive insufficient precipitation compared to reality. Whereas, if QPF is displaced in the north-south direction, the timing of when a flood wave may reach the gauge position can be greatly affected although the peak crest may not suffer serious errors.

Past studies have tried to quantify and account for spatial displacements. Carlberg et al. (2020) used a method for systematically shifting QPF to account for uncertainty caused by spatial displacements. In that study they used HRRRE as forcing by shifting each of its 9 original members, 8 times in all cardinal directions. The latitude and longitude shifts were made for both 0.5 degree and 1 degree, creating two 81 member ensembles when including the original HRRRE members. That ensemble QPF was used as forcing for the Hydrology Laboratory-Research Distributed Hydrologic Model (HL-RDHM). The HL-RDHM outputs showed that there was increased potential to detect flooding events. However, there were many forecast members that received zero-QPF, which led to low flood probabilities (Carlberg et al. 2020).

Kiel et al. (2021) set out to quantify and compare the climatological spatial displacements in the HREF and HRRRE. A threshold of 1 mm of accumulated precipitation was used to identify areas receiving rainfall. The precipitation regions of the QPFs and observations were compared using their centroids, or center of mass. The spatial displacements between the predicted and observed centroids were found for both accumulated 0 to 18-hour forecasts, and at the hour of convective initiation (CI). Across 30 cases, over the 2018 warm season, they found that for the CI displacements, the HRRRE had a smaller spread than the HREF (Kiel et al. 2021). It was also found that the HRRRE accumulated displacements had a systematic bias to the west, with an average magnitude across all members of 21 km (Kiel et al. 2021).

This work addresses the Office of Water Prediction's (OWP) mission statement of, "collaboratively research, develop and deliver timely and consistent, state-of-the-science national hydrologic analyses, forecast information, data, guidance, and decision-support services to inform essential emergency management and water resources decisions across all time scales" (OWP 2020), by experimenting with the data found in Kiel et al. (2021) to develop a more

refined methodology to reduce the number of zero-QPF members and improve streamflow ensemble forecasting over the same region of the Upper Mississippi Valley as used in previous hydrologic forecasting studies (Carlberg et al. 2020; Goenner et al. 2020; Kiel et al. 2021). Because climatological spatial displacement data for the HRRRE could be used from the work of Kiel et al. (2021), and methods for regridding HRRR inputs for the WRF-Hydro were available, the HRRRE was chosen as forcing for our streamflow ensembles. A method for randomly shifting QPF within the bounds of the climatological displacement identified by Kiel et al. (2021) was devised. The shifted QPF along with the original QPF were input into the NWM to generate ensembles streamflow predictions for 50 stream gauges, over 29 events from the 2018 warm season. Additionally, weighting schemes were created to refine the streamflow ensembles based on the position of the ensemble members from the calculated displacement at CI. Standard forecast verification methods were used to evaluate the streamflow ensembles for creation of both deterministic and probabilistic forecasts of peak streamflow.

3.3 Data and Methods

3.3.1 Models and Datasets

The nine-member High-Resolution Rapid Refresh Ensemble (HRRRE; Dowell et al. 2018; Dowell 2020) was the QPF product used in this study. It was chosen because of its utility as a CAE, as well as its use in the prior research done by Carlberg et al. (2020) to test QPF shifting, and Kiel et al. (2021) to quantify its spatial displacements. Perturbations of pressure, temperature, wind vectors U and V, water vapor mixing ratio fields are added to the HRRRE's boundary conditions to generate ensemble variety. Boundary conditions for the creation of the HRRRE are provided by the Rapid Refresh model, while data assimilation is accomplished using the National Oceanic and Atmospheric Association (NOAA) Gridpoint Statistical Interpolation analysis system (Benjamin et al. 2016). The land surface, planetary boundary layer, and

microphysics were handled by the Smirnova/Rapid Update Cycle, Mellor-Yamada-Nakanishi-Niino, and Thompson schemes, respectfully (Smirnova et al. 2016; Nakanishi and Niino 2009; Thompson and Eidhammer 2014). At the time data was compiled for the use in this study, the HRRRE's domain covered the eastern two-thirds of the contiguous United States (CONUS), with the model core being the HRRRv3 (Dowell et al. 2018). The experimental HRRRE forecasts used in this study were obtained from the NOAA Global Systems Laboratory using their ftp server.

Observed hydrologic model forcing came from the North American Land-Data Assimilation System version 2 (NLDAS-2; Xia et al. 2012a,b). NLDAS-2's domain covers the CONUS at 1/8 degree spacing. The North American Regional Reanalysis supplies several atmospheric data fields for use in NLDAS-2, including: 2 m air temperature, 2 m specific humidity, 10 m wind speed, precipitation, surface pressure, shortwave-, and longwave radiation. Observed data from the National Center for Environmental Prediction Stage-II doppler radar precipitation estimates and the Climate Prediction Center unified gauge-based precipitation data are disaggregated to the hourly timescale and assimilated onto the grid of NLDAS-2 (Xia et al. 2012a). NLDAS-2 data from October 2013 to April 2018 was used to create a 4 year spin-up for the hydrologic model. NLDAS-2 was also used in the creation of warm-starts for each HRRRE storm event.

The hydrologic model used in this work was the WRF-Hydro version 5.1.1, using the National Water Model 2.0 (NWM) configuration (Gochis et al. 2020). The land surface model used within the NWM is the Noah Multiparameterization (NoahMP), running on a 1 km horizontal grid, with a 2 m soil depth split into 4 layers, operating on hourly timesteps in accordance with the timestep of the forcing chosen for this study. Inputs to the NoahMP are

shown in Table 1. Data from the NoahMP is then used to generate the runoff of the modelled precipitation utilizing terrain routing on a 250 m nested grid. The Steepest Descent option was used for terrain routing, on 10 s timesteps. Groundwater was managed by the built-in exponential bucket model. Lastly, Muskingum-Cunge reach-based routing was used to convey water downstream after it entered the channel; the reach-based routing runs at a 300-s timestep. Streamflow is output at hourly timesteps for specified locations.

Observed streamflow needed for verification purposes was acquired from the United States Geological Survey (USGS) Waterdata database (USGS 2016). The data is available at 15-minute intervals and was averaged to one hour for this study. Missing timesteps were ignored and only timesteps with observed streamflow were used to eliminate skew from missing timesteps in the original fifteen-minute dataset.

3.3.2 Site and Case Selection

A cutout from the NWM that covers most of the state of Iowa, as well as portions of Southern Minnesota, Southwest Wisconsin, and Northwest Illinois, as shown in Fig. 1, was obtained from the National Center for Atmospheric Research (NCAR). A warm season period was used as forcing for WRF-Hydro streamflow to examine which gauges within the NWM cutout were able to accurately recreate observed flows, while also being able to identify gauges that were attached to incomplete watersheds. For example, the Mississippi River and Missouri River appear in the model domain, but their watersheds are not fully encompassed and were not accurately modelled. The WRF-Hydro was run from 1 May 2018 to 1 November 2018 using NLDAS-2 forcing. This period was chosen due to the occurrence of a wet summer in the study region, lending to high-flow events for analysis. While slightly below the thresholds deemed as satisfactory for model verification, a Nash-Sutcliffe Efficiency (NSE) greater than 0.4 and Percent Bias (Pbias) under 40% were used to identify gauges with passable performance

(Moriasi et al. 2007; Madsen et al. 2020; see appendix entry Figure A.1 for NSE and Pbias). As a result of the analysis, 50 gauges met the requirements and are displayed in Table 2.

Cases of interest for this work were identified when a flash flood warning, or watch, was issued by the National Weather Service (NWS), or if flooding was observed without prior advisory issuance (Table 3). The HRRRE forecasts were regridded using the HRRR regridding package provided by NCAR to isolate the NWM domain from the much larger HRRRE domain and scale the 3 km HRRRE data to the 1 km grid of the NoahMP.

3.3.3 Shifting QPF

In the research done by Kiel et al. (2021) the accumulated spatial displacements were calculated by comparing the position of the centroids of the accumulated rainfall of the precipitation systems provided in the QPFs and the observed rainfall for 30 cases from May 2018 to September 2018. While acknowledging the limitations of the short period of record covered by the Kiel et al. (2021) dataset, this study covers the same storm events, allowing for the Kiel et al. (2021) dataset to be considered a near perfect climatology. The spatial displacements of QPF for the HRRRE found in Kiel et al. (2021) followed a skewed normal distribution. A skewed-normal random number generator (snRNG) was used to select locations for random shifts of QPF within our model grid that fit the distributions noted above. The distributions for the latitude and longitude values were managed independently of each other when running the snRNG because, though they shared similarities in spread, their skew values had differing signs (shown in Fig. A.5 in the appendix).

To determine the “optimal” number of shifts per HRRRE member to use in this study, the accuracy of sample distributions produced by the snRNG were analyzed against the characteristics of the original distributions provided by Kiel et al. (2021), while also noting the amount of time needed to process and execute a full suite of model runs for ensembles of various

sizes. The smallest ensemble size that was analyzed had three shifts per parent HRRRE member, creating a 36-member ensemble (27 shifted members, and the 9 non-shifted members), while the largest ensemble had 8 shifts per member, creating an 81-member ensemble after including the non-shifted members, similar in size to those tested by Carlberg et al. (2020). The snRNG was run 100 times for each sampling regime: 36-, 45-, 54-, 63-, 72-, and 81-members, respectively. For each of the six sampling regimes, the mean absolute error was calculated for the mean, standard deviation, and skew, between the sample distribution and the climatological distribution provided by Kiel et al. (2021). As the number of shifts in the sampling regime increased, the error between the ensemble distribution and the climatological distribution generally decreased. However, the amount of time to preprocess the QPF and run the model for each case increased by about 25% as the sampling regime was increase in size by 9 shifted members and thus prohibited the use of the larger ensembles. To optimize both runtime and ensemble precision, 54 shifts were chosen. With the 9 non-shifted HRRRE members included, that meant the full ensemble would be populated by 63-members.

During the regridding process, data were mapped to the NWM domain, as well as a larger "shifting" domain that was meant to fully encompass the NWM domain with considerable buffer on every side, as shown in Figure 2. This was done so that the larger domain could act as a source of data for the shifted QPF forcing, while the smaller NWM domain provided the non-shifted dataset and a template to which data could be shifted into. As detailed in panel b) and c) of Figure 2, in the case of a northwest shift, data matching the grid dimensions of the NWM domain would be taken from the shifting domain and passed to the center of a copy of the NWM domain. This acted to deceive the model into thinking that precipitation that was originally to the northwest, is actually centered over the domain of study.

Calculations of QPF spatial displacement error at convective initiation (CI) were also provided by the work done by Kiel et al. (2021). Kiel et al. (2021) hypothesized that CI displacements could offer insight into the behavior the 0- to 18-hour accumulated displacement of the HRRRE QPFs in the forecast region. To build on that hypothesis, this study uses the CI displacements to refine the weighting of ensemble members to improve flood forecasting, while still leveraging the use of QPFs for increased lead time of flooding events over QPE products.

Kiel et al. (2021) had calculated these CI displacements for each model member and each event individually due to the differences in the members' behavior, displacement, and timing at CI. The second of the two weighting schemes adjusted magnitude and direction of the CI displacement using a correction, depending on what directional quadrant the CI displacement fell in (see figure A.3 in the appendix). This correction was brought about because Kiel et al. (2021) had diagnosed that spatial displacement errors at CI were larger than the accumulated 0- to 18-hour spatial displacement errors for forecast event.

Weights could be assigned to members by taking the inverse of the distance between the centroid of the precipitation systems in the shifted ensemble members and centroid of the observed spatial displacement of the precipitation systems at CI. Then the weighting values were normalized so that all ensemble member weights added up to 100%. Weights were also calculated for the non-shifted members to treat all members equally. There was an issue that arose if a model member had perfect placement of CI. If the displacement was 0 km, that would lead to that member being assigned an infinite weight. Instead, that member was given an arbitrary value of 0.5 km to avoid the division by 0, while still awarding that perfect displacement error with the highest possible weight. Because the weights are normalized by the total distances, the results are not sensitive to this change.

Two additional ensembles were added to our comparison by making use of CI displacement errors. In both of the new slimmed down ensembles, nine model members were isolated from the full 63-member dataset. These ensembles were intended to provide a direct comparison to the base HRRRE due to their identical ensemble size. The first of the two new 9-member ensembles was formed by first grouping all 54 shifted members with their parent HRRRE model member and then selecting the member from each parent HRRRE model member group with the smallest distance to the CI displacement (hereafter referred to as Closest A). An example of a parent model group is non-shifted member one and shifted members one through six. This ensemble was meant to keep the variability of each of the perturbation driven members of the base HRRRE, while isolating members that were closest to the observed behavior of the precipitation at CI.

The other new 9-member ensembles was formed by isolating the nine model members from the full 63-member ensemble that had the smallest distance to the CI displacement (hereafter referred to as Closest B). The Closest B was intended to see if the hydrologic prediction could be improved even with the chance of picking more than one member from a parent HRRRE member group. For both of the Closest A and B ensembles, once they were created, each member was given equal weight. To test the performance of the Kiel et al. (2021) correction of the CI displacement errors, this adjustment was tested in both of the Closest A and B ensembles. Because of their dependence on the use of CI displacement error in post-processing, the Closest A and B ensembles are only available for comparison when the two CI displacement weighting schemes are being used for the full ensemble and the base HRRRE. In all, four streamflow ensembles were evaluated. The full 63-member and original HRRRE ensembles being evaluated using equal weighting, CI displacement error weighting, and

corrected CI displacement error weighting. When the CI displacement errors are being used in the weighting schemes for the full ensemble and base HRRRE members, the Closest A and B ensembles will be evaluated alongside them with the use of the Kiel et al. (2021) correction factored in.

3.3.4 Verification

Deterministic and probabilistic forecast skill of the magnitude of peak discharge was evaluated using several standard forecast verification metrics. Frequency of non-exceedance (FNE) assesses whether a forecast was able to capture the observation within the bounds of the ensemble (Carlberg et al. 2020):

$$FNE = \frac{\sum_{t=1}^{N} I[y_t]}{N} \quad (1)$$

$$I[y_t] = \begin{cases} 1, L_t < y_t < U_t \\ 0, otherwise \end{cases}, \quad (2)$$

where L_t and U_t are the lower and upper bounds of the ensemble, I is the Boolean value pertaining to an individual forecast of y_t, and N is the total number of forecasts. Because of this study's focus on evaluating flood forecast, the ensemble was not given a lower bound, thus setting L_t to zero streamflow, this was so that the ensemble should not be penalized for errors in baseflow. A perfect FNE is a score of 1, meaning the events were all contained by the ensemble forecasts, with the worst possible score being 0. It is important to note that FNE is not affected by member weighting, as the weight is independent of the member forecast. Whereas, for the Closest A and B ensembles, FNE could change as different model members were selected.

The ranked probability score (RPS) for any individual forecast is the sum of the squared differences of the cumulative distribution of the forecast (F_m) and the observed (O_m) (Wilks 1995; Franz et al. 2003):

$$RPS = \sum_{m=i}^{J} (F_m - O_m)^2 \qquad (3)$$

The cumulative distribution of the forecast (F_m) is:

$$F_m = \sum_{j=1}^{m} f_j\,, \qquad m = 1, \ldots, J, \qquad (4)$$

where f_j is the cumulative probability of the forecast, J is the total number of forecast categories, and m is the forecast category (Wilks 1995). The cumulative distribution of the observed streamflow (O_m) is:

$$O_m = \sum_{j=i}^{m} o_j\,, \qquad m = 1, \ldots, J \qquad (5)$$

Where the category in which the observed discharge peak, o_j, occurs is given a value of 1, and all categories greater than o_j also receiving a value of 1. An RPS value of 0 is ideal. Smaller scores mean that there is less difference between the forecast probability and the observed probability. The forecast categories used were less than 50% of action stage, 50% of action stage to action stage, action stage to minor stage, minor stage to moderate stage, moderate stage to major stage, and greater than major stage. The values for each category for each gage were found at the NCRFC website. Because RPS is dependent on the gauging point having a major stage defined by the NCRFC, this metric could only be calculated at 37 of the 50 gauges.

A 2x2 contingency table was used to evaluate the forecast skill for prediction of flood or no flood where "flood" is defined as minor flood stage (Table 4). Of our original 50 gauges that met our standards for NSE and Pbias, only 43 reported having a minor stage. If flooding was observed, and the model predicted flooding, the result is a hit (H). Otherwise, if flooding was not

predicted by the model and flooding was observed, it is assigned a miss (M). Similarly, when flooding was not observed and the model predicted flooding, it is assigned a false alarm (FA). If flooding was not observed and the model supported that finding, it is given the distinction of a correct negative (CN).

There are several forecast metrics that can be calculated using the information in the 2x2 contingency table. Probability of detection (POD), the ratio of the number of hits to the total number of events observed:

$$POD = \frac{H}{H + M}. \tag{6}$$

False alarm ratio (FAR) is the ratio of the number of false alarms FA to the total number of events forecasted to occur:

$$FAR = \frac{FA}{FA + H}. \tag{7}$$

Critical success index (CSI) is the ratio of hits to all observed and forecasted events:

$$CSI = \frac{H}{H + FA + M}. \tag{8}$$

Equitable threat score (ETS; also called Gilbert skill score) is quite similar to CSI, however it adds an estimation for chance (Equation 10) to correct for the number of hits that may have occurred due to a chance forecast. The full expression for ETS is given as:

$$ETS = \frac{H - Chance}{H + FA + M - Chance}. \tag{9}$$

Chance is calculated as the events forecasted multiplied by the events observed divided by the total number of forecasts N:

$$Chance = \frac{(H + FA) * (H + M)}{N}. \tag{10}$$

Finally, the Heidke skill score (HSS; Heidke 1926) also attempts to account for chance, while also taking into account the number of correct negatives being produced by the model (See Equation 12).

$$HSS = 2 * \frac{(H * CN) - (FA * M)}{((H + M) * (M + CN) * (H + FA) * (FA + CN))} \quad (11)$$

For both ETS and HSS, a value less than 0 means the model did worse than what would be expected from a chance forecast, with the best outcome being a value of 1.

Probability of exceedance (POE) thresholds were used to form several contingency tables in order to assess ensemble forecasting at differing levels of ensemble certainty. POE can be described as the accumulated weight of model members that had forecasted peak streamflow at or above minor flood stage. The following POE thresholds were used: >0%, ≥10%, ≥20%, ≥30%, ≥40%, ≥50%, ≥60%, ≥70%, ≥80%, and ≥90%.

In addition to the probabilistic measures mentioned previously, reliability was calculated using forecast probability intervals of 0-5%, 5-15%, 15-25%, 25-35%, 35-45%, 45-55%, 55-65%, 65-75%, 75-85%, 85-95%, and 95-100%. The reliability is displayed as a diagram which represents the conditional distribution of observed events given the forecast of an event (Wilks 1995), in this case pertaining to the exceedance of minor flood stage. Perfect reliability occurs when the data lies along the 1:1 line across the diagram, forecasts are over-forecasting when they fall to the right of the 1:1 line, and under forecasting when they fall to the left of the 1:1 line (see figure A.2 in the Appendix). As a numerical assessment of reliability, the mean absolute error (MAE) is calculated (Wilks 1995):

$$MAE = \frac{\sum_{i=1}^{N} |(F_i - O_i)|}{N}, \quad (12)$$

where the relative frequency of the observations, O_i, is subtracted from forecasted probability, F_i, for all forecast probability intervals (i…N).

Though the focus of this research was meant to look at ensemble performance based on the magnitude of peak discharge, it also briefly examines the timing of peak discharge and any disparity in forecast skill for large and small watersheds. Seven categories were used to evaluate the timing of the predicted peak discharge: more than 6 hours early, 4-6 hours early, 1-3 hours early, timing hits (hour 0), 1-3 hours late, 4-6 hours late, and greater than 6 hours late. The minimum threshold for flood peak used in this analysis was action stage, therefore, any member that did not produce output above action stage were disregarded in the timing analysis.

To investigate the difference between different sized basins, a threshold had to be set to define a "large" and "small" basin. Of the gauges that experienced flooding across all forecast events, the median watershed area was 3437 km^2, with a mean area of 5531 km^2, resulting in a decision of 4000 km^2 to be used as the cutoff. All metrics, including probabilistic and deterministic measures, as well as the distributions of member peak discharge timing were compared across that 4000 km^2 boundary using a statistical test. Moreover, all p-values for determining statistical significance, for all metrics, were gathered using a paired two-tailed T-test. A 90% confidence interval was used as the threshold for assessing significance. These tests were applied to pairs of values linked by the same event, when applicable.

3.4 Results

3.4.1 Probabilistic Measures

The FNE of the full ensemble of 63 members showed better performance than the base HRRRE members by roughly 11%, visualized in Figure 3a, with specific values shown in Table 6. The results for the Closest A and B ensembles showed that they had a lower FNE than the full ensemble, while still displaying improved FNE over the base HRRRE members. The Closest A

ensemble had better FNE values than the Closest B, though this difference was only statistically significant when using corrected CI displacements. Overall, both of the Closest A and B ensembles showed statistically significant improved performance over the base HRRRE, though they did fall short of the full ensemble. The full ensemble had significantly better performance in FNE compared to the other three ensembles no matter the weighting scheme (Table 7).

For RPS, the full ensemble produced more accurate forecasts as compared to the base HRRRE. RPS values for both the base HRRRE and full ensemble got worse when the weighting schemes were used, with the poorest performance occurring while using the Kiel et al. (2021) corrections, shown in Fig. 3b. The Closest A and B ensembles showed little change between the weighting schemes. Like the results of the t-tests found for FNE, the Closest A and B ensembles were significantly better than the base HRRRE for RPS, with the full ensemble being statistically better than the other three ensembles at our 90% confidence interval as shown in Table 7.

In Figure 4, the reliability curves are shown for all the ensembles and all weighting schemes. The gap at 50% reliability for the base HRRRE is due to there only being nine members in this ensemble. For the base HRRRE, POE values are possible at 44.4% and 55.5% but do not lie within the 45% to 55% POE category. That is why reliability was unavailable for the category at 50%. Comparison of the reliability of all four ensembles using the CI displacement weighting schemes is harder to visually evaluate. The MAE data provided in Table 8 allow for easier analysis of these differences. The best value for MAE was produced from the full ensemble when equal weights were used. The full ensemble consistently had lower MAE values than the base HRRRE. The use of CI displacements and corrected CI displacements decreased the reliability of the full ensemble, with both schemes producing a MAE of 0.174. By averaging across all three weighting schemes, there was a mean increase in reliability for the full

ensemble of 28% over the base HRRRE. The worst overall score was a MAE of 0.234, which was produced by the base HRRRE using CI displacement weighting. The Closest A and B ensembles generally produced values in between the full ensemble and the base HRRRE. The Closest B was statistically better than the base HRRRE at the 90% confidence interval. It should be noted that all the ensembles, no matter what weighting was used, saw POD values lower than the observed frequency of flooding.

3.4.2 Deterministic Measures

For the deterministic verification, ten thresholds for POE made it possible to evaluate ensemble accuracy and ensemble certainty, where higher POE thresholds correspond to higher ensemble certainty. Therefore, values in higher POE thresholds came about from greater numbers of ensemble members having forecasted minor flooding. Skill metrics were calculated at each POE threshold to determine at which threshold peak model performance was occurring. This threshold of peak performance helped to offer insight into ensemble certainty. The results showing all values at all POE thresholds for equal weighting are shown in Table 9. Bolded values in Table 9 represent the peak values of the skill metrics for each ensemble. It should be noted that because the base HRRRE only contains nine members, the lowest possible POE is already greater than 10% ($\frac{1}{9} \sim 11\%$) for equal weighting. As can be seen from the data, the magnitude of CSI, ETS, and HSS produced by the full ensemble were greater than the base HRRRE. Those scores were 0.472, 0.419, and 0.591, respectively. However, the base HRRRE reaches its best values for CSI, ETS, and HSS at a higher certainty, a threshold of 70% compared to the 60% threshold of the full ensemble. Looking at the p-values provided in Table 12, the differences between the full ensemble and the base HRRRE for POD were the only statistically

significant values at a 90% confidence interval, though the adjacent p-value for FAR and CSI would have been significant at an 85% confidence interval.

When the weighting scheme was used to account for CI displacement error, the threshold that saw peak skill scores for the full ensemble moved up to 70%, implying that it had roughly equal forecasting certainty to the base HRRRE (Table 10). There were increases in skill metrics between the equal weighting and CI displacement weighting. The peak values for CSI went from 0.472 to 0.478, ETS increased from 0.419 to 0.430, and HSS improved from 0.591 to 0.601 for the full ensemble. For the base HRRRE, its peak value for CSI was once again at the 70% threshold; however, its peak values for ETS and HSS moved up to the 90% threshold. This was the only time that the peak values of the skill measures did not all occur at the same POE threshold for an ensemble.

Using the CI displacement errors allowed for the creation and examination of the two slimmed down Closest A and B ensembles. The Closest A and B ensembles both saw their maxima in skill scores around the 60% threshold. The p-values do not support definitive differences for the CI displacement error weighting scheme with only the POD of the full ensemble being significantly lower than the base HRRRE at the 90% confidence interval.

When the Kiel et al. (2021) corrections were applied to the CI displacement errors, we see that the peak skill values increased for the base HRRRE, Closest A, and Closest B over their values produced while using the non-corrected weights. The full ensemble saw near equal or lower peak values to what it produced using the CI weighting scheme (Table 11). Interestingly, the full ensemble peak skill values moved from the 70% threshold, produced using the non-corrected weights, to the 50% threshold for the corrected weights. This is the lowest threshold to represent the peak of any ensemble using any of the weighting schemes. This jump between POE

thresholds was likely caused by a decrease in FAR, though this still calls into question the certainty in this ensemble's forecasting ability using the corrected weighting. The Closest A saw a slight improvement in forecasting certainty by shifting its peak from 60% to 70%. Overall, comparisons of the deterministic skill metrics of the ensembles at our POE thresholds yielded little statistical significance. The only two significant comparisons came between the base HRRRE and the Closest A for CSI and ETS using the corrected CI displacement weights, shown in Table 12.

3.4.3 Flood Peak Timing and Watershed Drainage Area Analysis

Table 13 shows how member weights were distributed using timing categories of greater than 6 hours early, 4-6 hours early, 1-3 hours early, timing hits (hour 0), 1-3 hours late, 4-6 hours late, and greater than 6 hours late. All the ensembles in this study struggled to forecast flood wave peaks within a 6-hour window of the observed peak timing. The highest cumulative percentage of members between 6 hours early and 6 hours late was produced by the Closest A using the uncorrected CI displacement errors, with 23.44%. On average over 78% of ensemble members forecasted peak discharge to occur more than 6 hours early or 6 hours late. There were no statistically significant differences between the distributions of peak timing.

Comparisons of ensemble performance based on watershed area showed that often, larger watersheds had higher deterministic skill values at all POE thresholds. For all three weighting schemes, the better skill values produced by watersheds with greater than 4000 km^2 were statistically significant compared to the smaller watersheds. The same relationship was seen for RPS, with the ensembles producing peak forecasts that more closely matched observations for the larger basins. On the other hand, FNE showed that the smaller watersheds were more easily contained by the ensemble forecasted maximum discharge, with higher FNE outputs for nearly all ensembles and weights. The only model to see a better FNE for larger basins was the Closest

A using CI displacement weighting. All differences found between watersheds larger and smaller than 4000 km^2, for FNE and RPS, were found to be statistically significant at a 90% confidence interval. This is unsurprising given that the scale of the smaller watersheds is close to if not smaller than the scale of the rain area of these intense convective systems, allowing variations and displacement to contribute large errors for these smaller basins.

When comparing larger and smaller watersheds using flood peak timing data, there were no clear distinctions between the distributions of member weights. This was the case for all weighting schemes. The t-tests support the parity between the distributions with p-values of 1 for all three weighting schemes, leading to no statistical significance.

3.5 Discussion

It was found that the full ensemble had higher peak values for the deterministic skill measures when using equal weighting and uncorrected CI displacement weights, which is similar to what was found by Carlberg et al. (2020). However, when the corrected CI displacement weights were used, our study found that the base HRRRE produced the best overall peak skill values for CSI, ETS, and HSS. The POE thresholds that demonstrated the highest skill were always at or above 50%. As a point of contrast, Reed and MacFarlane (2020) used a 30% exceedance threshold to denote minor risk and a 70% threshold for high risk for ensemble forecasts. Our data suggest that lower thresholds may not be as relevant for ensemble forecasts produced by HRRRE forcing.

The gains and losses across the different ensembles were generally small, with a nearly universal lack of statistical significance for the deterministic measures. The full ensemble significantly outperformed, at a 90% confidence level, the other three ensembles in terms of both FNE and RPS, no matter which weighting scheme was used. Due to this fact, it appears that the informed spatial shifting of QPF from the HRRRE produced improved probabilistic forecasts,

and relatively equal deterministic forecasting ability as compared to the non-shifted original HRRRE members. The two Closest A and B ensembles performed better in the probabilistic skill measures than the base HRRRE, despite still being worse than the full ensemble. For deterministic skill, the Closest B had the best score for the CI displacement weighted CSI, but otherwise they fell short of either the full ensemble and/or the base HRRRE. There were no instances of statistical significance to support them being a better choice compared to the full ensemble and/or the base HRRRE.

For the comparison of the different weighting schemes, although there were a handful of statistically significant differences present for POD and FAR, there were no significantly significant differences for CSI, ETS, or HSS. For reliability, the equal weighting produced the lowest MAE, with the full ensemble performing the best. The CI displacement weighting schemes increased the MAE by about 16% for the full ensemble, while the base HRRRE produced the worst MAE score when the non-corrected CI displacement error weighting scheme was used. However, none of the changes in reliability between schemes were significant at a 90% confidence interval. For FNE, the increase seen for the Closest B when using the corrected CI displacement errors was significant over the FNE when using uncorrected CI displacement errors. Lastly, RPS values worsened when the base HRRRE was adjusted to use the two CI displacement weighting schemes. Those differences were statistically significant at our 90% confidence interval. Similarly, the decrease in RPS for the full ensemble was also statistically significant when using the two CI displacement weighting schemes. Differences in the Closest A and B ensembles were not statistically significant between the corrected and uncorrected CI displacement error selection processes.

The correction used in the third weighting scheme was meant to reduce the drastic displacements that may be seen at CI to recreate the accumulated QPF displacements seen through forecast hour 18. The correction improved some measures for the base HRRRE, but more often than not, reduced skill values for the other three ensembles. Although the rain events used in this study had most of the precipitation falling within the first 18 hours of the forecast period to match the 0- to 18-hour climatological spatial displacements found in Kiel et al. (2021), our QPFs were still allowed to run out to the end of the 36-hour HRRRE forecast. The correction was meant to reduce the value of the observed spatial displacement at CI to more closely match the behaviors of the 0- to 18-hour accumulated spatial displacement. Those reductions may be less relevant as the displacement accumulation continues through forecast hour 36. Thus, more research needs to be done to test the validity of the Kiel et al. (2021) correction for CI displacements for longer forecast periods.

The comparisons of “large” and “small” watersheds in this study support the idea that larger basins and the responses they display were more easily modelled within this system, which further matches previous research (Merz et al. 2009; van Esse et al. 2013; Poncelet et al. 2017). This is likely due to large watersheds being less sensitive to the displacement errors of the precipitation systems, than the basins with smaller areal coverage. The large basins in this study had significantly higher scores of CSI, ETS, and HSS at all POE thresholds at a 90% confidence interval. For probabilistic measures, large basins produced better RPS values. FNE was the only metric that produced better scores for small basins. The model may be over estimating runoff for these isolated storms at times, helping to increase the FNE in the smaller basins that are more prone to flash flooding. However, that speculation would need to be investigated further. Both findings for RPS and FNE mentioned above were also statistically significant.

The analysis for the timing of peak discharge showed that all four ensembles struggled to match observations for these flood events. Over three quarters of the ensemble weights were over 6 hours early or 6 hours late. The shifting of QPF appears to have not effected the timing of peak discharge, as there were only minor differences between the ensembles with no statistical significance. Given the data found here, it cannot be determined whether the temporal errors were contributed by the WRF-Hydro, HRRRE QPF, or some combination of both. Additional work should refine the shifting methods further to account for temporal errors in these hydrologic simulations.

3.6 Summary and Conclusions

This research aimed to test if implementing randomized spatial shifting of QPF forcing driven by a climatological spatial displacement dataset could improve ensemble streamflow forecasts for flood events in a region of the Upper Mississippi Valley. To do so, a skewed normal random number generator was used to produce shifted QPF members to fall within the 0- to 18-hour accumulated spatial displacement distributions found by Kiel et al. (2021) for the HRRRE. In total, four ensembles: the full 63-member ensemble, the original HRRRE, the 9-member ensemble with selection based on the member with the lowest CI displacement error from each original parent HRRRE member group (Closest A), and the 9-member ensemble with member selection based on the shortest distance to observed spatial displacement at CI (Closest B), were compared. Those comparisons were done using three weighting schemes: equal weighting, CI displacement weighting, and corrected CI displacement weighting where applicable. The magnitude of forecast peak discharge was evaluated using both probabilistic and deterministic skill measures. Ensemble performance for the timing of peak discharge and ensemble prediction for small and large watersheds were also briefly examined. This analysis was carried out using

50 stream gauges over 29 cases from the 2018 warm season, with statistical significance of differences determined using paired two-tailed T-tests at a 90% confidence interval.

This research found that the informed shifting of QPF to create ensemble members for the purposes of generating streamflow forecasts led to near equal performance when looking at the deterministic skill measures of CSI, ETS, and HSS at several POE thresholds. However, the ensembles made up of shifted members showed greatly improved probabilistic forecast skill using FNE, RPS, and reliability.

The addition of the CI displacement weighting schemes showed mixed results. The uncorrected weights showed some increased skill over equal weighting of ensemble members, while the corrected CI displacements produced marginally lower scores. Moreover, the Closest A and B ensembles usually had a performance that fell between the full ensemble and the base HRRRE. Although the weighting schemes did not show significant improvement at our 90% confidence interval, further refinement of the weighting schemes may yield more usefulness for the implementation of the CI displacements as well as the Closest A and B ensembles.

Timing of peak discharge appears to be the greatest shortcoming of the forecast systems created in this study. Neither informed shifting nor CI displacement weighting produced differences in the forecasted timing of flood peaks. The comparison of ensemble performance for larger and smaller watersheds found that forecasts for larger basins were much more skillful for every metric besides FNE.

3.7 Acknowledgements

Funding for this work was provided through the National Oceanic and Atmospheric Administration Collaborative Science Technology, and Applied Research grant NA17NWS4680005. We would like to thank the NCAR WRF-Hydro team for their unending help in the troubleshooting of model issues, and for their assistance in creating the NWM cutout.

The author would like to thank the previous researchers who worked on this grant and provided the foundation for this research, Bradley Carlberg, Andrew Goenner, and Benjamin Kiel. Thanks go out to Jonathan Thielen for help with python troubleshooting and Elizabeth Tirone for help editing.

3.8 References

Adams, T. E., 2016: Flood Forecasting in the United States NOAA/National Weather Service. *Flood Forecasting: A Global Perspective*, Elsevier Inc., 249–310.

——, and R. Dymond, 2019: The effect of QPF on real-time deterministic hydrologic forecast uncertainty. *J. Hydrometeorol.*, **20**, 1687–1705, https://doi.org/10.1175/JHM-D-18-0202.1.

American Meteorological Society, 2012: Glossary of Meteorology. *Am. Meteorol. Soc.*,. https://glossary.ametsoc.org/wiki/Mesoscale_convective_system (Accessed October 12, 2020).

Andresen, J., S. Hilberg, K. Kunkel, J. Winkler, J. Andresen, J. Hatfield, D. Bidwell, and D. Brown, 2012: *Historical Climate and Climate Trends in the Midwestern USA National Climate Assessment Midwest Technical Input Report*. http://glisa.msu.edu/docs/NCA/MTIT_Historical.pdf. (Accessed December 9, 2020).

Benjamin, S. G., and Coauthors, 2016: A North American hourly assimilation and model forecast cycle: The rapid refresh. *Mon. Weather Rev.*, **144**, 1669–1694, https://doi.org/10.1175/MWR-D-15-0242.1.

Berenguer, M., M. Surcel, I. Zawadzki, M. Xue, and F. Kong, 2012: The diurnal cycle of precipitation from continental radar mosaics and numerical weather prediction models. Part II: Intercomparison among numerical models and with Nowcasting. *Mon. Weather Rev.*, **140**, 2689–2705, https://doi.org/10.1175/MWR-D-11-00181.1.

Brown, J. D., and G. B. M. Heuvelink, 2005: Assessing Uncertainty Propagation through Physically Based Models of Soil Water Flow and Solute Transport. *Encyclopedia of Hydrological Sciences*, John Wiley & Sons, Ltd.

Carlberg, B., K. J. Franz, and W. A. J. Gallus, 2020: A Method to Account for QPF Spatial Displacement Errors in Short-Term Ensemble Streamflow Forecasting. *Water*, **12**, 3505, https://doi.org/10.3390/w12123505.

Clark, A. J., W. A. Gallus, M. Xue, and F. Kong, 2009: A comparison of precipitation forecast skill between small convection-allowing and large convection-parameterizing ensembles. *Weather Forecast.*, **24**, 1121–1140, https://doi.org/10.1175/2009WAF2222222.1.

Collier, C. G., 2007: Flash flood forecasting: What are the limits of predictability? *Q. J. R. Meteorol. Soc. Q. J. R. Meteorol. Soc*, **133**, 3–23, https://doi.org/10.1002/qj.29.

Dowell, D., 2020: *HRRR Data-Assimilation System (HRRRDAS) and HRRRE Forecasts*. 1–8 pp. https://rapidrefresh.noaa.gov/internal/pdfs/2020_Spring_Experiment_HRRRE_Documentation.pdf.

——, C. Alexander, T. Alcott, and T. Ladwig, 2018: *HRRR Ensemble (HRRRE) Guidance 2018 HWT Spring Experiment*. 1–6 pp. https://rapidrefresh.noaa.gov/internal/pdfs/2018_Spring_Experiment_HRRRE_Documentation.pdf.

Du, J., S. L. Mullen, and F. Sanders, 1997: Short-range ensemble forecasting of quantitative precipitation. *Mon. Weather Rev.*, **125**, 2427–2459, https://doi.org/10.1175/1520-0493(1997)125<2427:SREFOQ>2.0.CO;2.

Ebert, E. E., 2001: Ability of a poor man's ensemble to predict the probability and distribution of precipitation. *Mon. Weather Rev.*, **129**, 2461–2480, https://doi.org/10.1175/1520-0493(2001)129<2461:AOAPMS>2.0.CO;2.

van Esse, W. R., C. Perrin, M. J. Booij, D. C. M. Augustijn, F. Fenicia, D. Kavetski, and F. Lobligeois, 2013: The influence of conceptual model structure on model performance: a comparative study for 237 French catchments. *Hydrol. Earth Syst. Sci.*, **17**, 4227–4239, https://doi.org/10.5194/hess-17-4227-2013.

Franz, K. J., H. C. Hartmann, S. Sorooshian, and R. Bales, 2003: Verification of National Weather Service Ensemble Streamflow Predictions for water supply forecasting in the Colorado River Basin. *J. Hydrometeorol.*, **4**, 1105–1118, https://doi.org/10.1175/1525-7541(2003)004<1105:VONWSE>2.0.CO;2.

Fritsch, J. M., R. J. Kane, and C. R. Chelius, 1986: The contribution of mesoscale convective weather systems to the warm-season precipitation in the United States. *J. Clim. Appl. Meteorol.*, **25**, 1333–1345, https://doi.org/10.1175/1520-0450(1986)025<1333:TCOMCW>2.0.CO;2.

Gallus, W. A., 2010: Application of object-based verification techniques to ensemble precipitation forecasts. *Weather Forecast.*, **25**, 144–158, https://doi.org/10.1175/2009WAF2222274.1.

Gallus, W. A. J., 2012: The Challenge of Warm-Season Convective Precipitation Forecasting. *Rainfall Forecasting*, Nova Science Publishers, 129–160.

Gochis, D. J., and Coauthors, 2020: *The NCAR WRF-Hydro® Modeling System Technical Description*.

Goenner, A. R., K. J. Franz, W. A. J. Gallus, and B. Roberts, 2020: Evaluation of an Application of Probabilistic Quantitative Precipitation Forecasts for Flood Forecasting. *Water*, **12**, 2860, https://doi.org/10.3390/w12102860.

Haberlie, A. M., and W. S. Ashley, 2019: A radar-based climatology of mesoscale convective systems in the United States. *J. Clim.*, **32**, 1591–1606, https://doi.org/10.1175/JCLI-D-18-0559.1.

Heidke, P., 1926: Berechnung des Erfolges und der Güte der Windstärkevorhersagen im SturmwarnungsdienstBerechnung des Erfolges und der Gute der Windstarkevorhersagen im Sturmwarnungsdienst. *Geogr. Ann.*, **8**, 301, https://doi.org/10.2307/519729.

Hejazi, M. I., and M. Markus, 2009: Impacts of Urbanization and Climate Variability on Floods in Northeastern Illinois. *J. Hydrol. Eng.*, **14**, 606–616, https://doi.org/10.1061/(asce)he.1943-5584.0000020.

Iyer, E. R., A. J. Clark, M. Xue, and F. Kong, 2016: A comparison of 36-60-h precipitation forecasts from convection-allowing and convection-parameterizing ensembles. *Weather Forecast.*, **31**, 647–661, https://doi.org/10.1175/WAF-D-15-0143.1.

Kain, J. S., S. J. Weiss, J. J. LevIt, M. E. Baldwin, and D. R. Bright, 2006: Examination of convection-allowing configurations of the WRF model for the prediction of severe convective weather: The SPC/NSSL Spring Program 2004. *Weather Forecast.*, **21**, 167–181, https://doi.org/10.1175/WAF906.1.

Kiel, B. M., W. A. J. Gallus, and K. J. Franz, 2021: A Climatology of Precipitation Displacement Errors in High-Resolution Ensembles. *J. Hydrometeorol.*,.

Lin, C., S. Vasić, A. Kilambi, B. Turner, and I. Zawadzki, 2005: Precipitation forecast skill of numerical weather prediction models and radar nowcasts. *Geophys. Res. Lett.*, **32**, n/a-n/a, https://doi.org/10.1029/2005GL023451.

Madsen, T., K. Franz, and T. Hogue, 2020: Evaluation of a Distributed Streamflow Forecast Model at Multiple Watershed Scales. *Water*, **12**, https://doi.org/10.3390/w12051279.

Merz, R., J. Parajka, and G. Blöschl, 2009: Scale effects in conceptual hydrological modeling. *Water Resour. Res.*, **45**, https://doi.org/10.1029/2009WR007872.

Moriasi, D. N., J. G. Arnold, M. W. Van Liew, R. L. Bingner, R. D. Harmel, and T. L. Veith, 2007: Model Evaluation Guidelines for Systematic Quantification of Accuracy in Watershed Simulations. *Trans. ASABE*, **50**, 885–900, https://doi.org/10.13031/2013.23153.

Nakanishi, M., and H. Niino, 2009: Development of an Improved Turbulence Closure Model for the Atmospheric Boundary Layer. *J. Meteorol. Soc. Japan*, **87**, 895–912, https://doi.org/10.2151/jmsj.87.895.

Nash, J. E., and J. V. Sutcliffe, 1970: River flow forecasting through conceptual models part I - A discussion of principles. *J. Hydrol.*, **10**, 282–290, https://doi.org/10.1016/0022-1694(70)90255-6.

National Weather Service, 2019: Weather Related Fatality and Injury Statistics. *Natl. Ocean. Atmos. Assoc.*,. https://www.weather.gov/media/hazstat/80years.pdf.

Office of Water Prediction, 2020: Welcome to the Office of Water Prediction. *Natl. Ocean. Atmos. Assoc.*,. https://water.noaa.gov/.

Poncelet, C., R. Merz, B. Merz, J. Parajka, L. Oudin, V. Andréassian, and C. Perrin, 2017: Process-based interpretation of conceptual hydrological model performance using a multinational catchment set. *Water Resour. Res.*, **53**, 7247–7268, https://doi.org/10.1002/2016WR019991.

Pryor, S. C., J. A. Howe, and K. E. Kunkel, 2009: How spatially coherent and statistically robust are temporal changes in extreme precipitation in the contiguous USA? *Int. J. Climatol.*, **29**, 31–45, https://doi.org/10.1002/joc.1696.

Reed, S. M., and A. MacFarlane, 2020: Validation of NWS Hydrologic Ensemble Forecast Service (HEFS) Real-Time Products at the Middle Atlantic River Forecast Center. *34th Conference on Hydrology*, Boston, MA https://ams.confex.com/ams/2020Annual/webprogram/Paper363657.html.

Roberts, B., B. T. Gallo, I. L. Jirak, A. J. Clark, D. C. Dowell, X. Wang, and Y. Wang, 2020: What does a convection-allowing ensemble of opportunity buy us in forecasting thunderstorms? *Weather Forecast.*, 1–65, https://doi.org/10.1175/WAF-D-20-0069.1.

Seo, B. C., F. Quintero, and W. F. Krajewski, 2018: High-resolution QPF uncertainty and its implications for flood prediction: A case study for the eastern Iowa flood of 2016. *J. Hydrometeorol.*, **19**, 1289–1304, https://doi.org/10.1175/JHM-D-18-0046.1.

Smirnova, T. G., J. M. Brown, S. G. Benjamin, and J. S. Kenyon, 2016: Modifications to the Rapid Update Cycle land surface model (RUC LSM) available in the weather research and forecasting (WRF) model. *Mon. Weather Rev.*, **144**, 1851–1865, https://doi.org/10.1175/MWR-D-15-0198.1.

Thompson, G., and T. Eidhammer, 2014: A study of aerosol impacts on clouds and precipitation development in a large winter cyclone. *J. Atmos. Sci.*, **71**, 3636–3658, https://doi.org/10.1175/JAS-D-13-0305.1.

U.S. Geological Survey, 2016: National Water Information System data available on the World Wide Web (USGS Water Data for the Nation). http://waterdata.usgs.gov/nwis/ (Accessed March 24, 2020).

Vasiloff, S. V., and Coauthors, 2007: Improving QPE and Very Short Term QPF: An Initiative for a Community-Wide Integrated Approach in: Bulletin of the American Meteorological Society Volume 88 Issue 12 (2007). *Bull. Am. Meteorol. Soc.*, **88**, 1899–1911, https://doi.org/10.1175/BAMS-88-12-1899.

Viterbo, F., and Coauthors, 2020: A multiscale, hydrometeorological forecast evaluation of national water model forecasts of the may 2018 Ellicott City, Maryland, Flood. *J. Hydrometeorol.*, **21**, 475–499, https://doi.org/10.1175/JHM-D-19-0125.1.

Wilks, D. S., 1995: Statistical Methods in Atmospheric Sciences. *Statistical Methods in Atmospheric Sciences*, 233–283.

Xia, Y., and Coauthors, 2012a: Continental-scale water and energy flux analysis and validation for the North American Land Data Assimilation System project phase 2 (NLDAS-2): 1. Intercomparison and application of model products. *J. Geophys. Res. Atmos.*, **117**, https://doi.org/10.1029/2011JD016048.

——, and Coauthors, 2012b: Continental-scale water and energy flux analysis and validation for North American Land Data Assimilation System project phase 2 (NLDAS-2): 2. Validation of model-simulated streamflow. *J. Geophys. Res. Atmos.*, **117**, https://doi.org/10.1029/2011JD016051.

Yan, H., and W. A. Gallus, 2016: An evaluation of QPF from the WRF, NAM, and GFS models using multiple verification methods over a small domain. *Weather Forecast.*, **31**, 1363–1379, https://doi.org/10.1175/WAF-D-16-0020.1.

Zhang, J., Y. Qi, C. Langston, B. Kaney, and K. Howard, 2014: A real-time algorithm for merging radar QPEs with rain gauge observations and orographic precipitation climatology. *J. Hydrometeorol.*, **15**, 1794–1809, https://doi.org/10.1175/JHM-D-13-0163.1.

3.9 Tables

Table 1: Fields needed from the source of forcing to run the NoahMP LSM in WRF-Hydro.

Variable	Units
Incoming shortwave radiation	(W/m^2)
Incoming longwave radiation	(W/m^2)
Specific humidity	(kg/kg)
Air temperature	(K)
Surface pressure	(Pa)
Near surface wind (u-component)	(m/s)
Near surface wind (v-component)	(m/s)
Liquid water precipitation rate	(mm/s)

Table 2: All gauge locations used in this study as well as watershed areas in square kilometers, and the maximum flood stage assigned to each gauge.

Station	**Watershed area (km^2)**	**Max Stage**
Floyd River at Alton, IA	694	Major
Little Sioux River at Linn Grove, IA	4009	Major
Little Sioux River at Correctionville, IA	6475	Major
Little Sioux River near Turin, IA	9132	Major
Wapsipinicon River near De Witt, IA	6050	Major
South Fork Iowa River NE of New Providence, IA	580	Action
Boone River near Goldfield, IA	1083	NA
Clear Creek near Coralville, IA	254.1	Minor
South Skunk River below Squaw Creek near Ames, IA	1440	Major
Iowa River near Rowan, IA	1111	Major
Indian Creek near Mingo, IA	715	Action
Maquoketa River at Manchester, IA	712	Major
Pecatonica River at Darlington, WI	707	Major
Boone River nr Webster City, IA	2186	Major
Old Mans Creek near Iowa City, IA	521	Minor
Winnebago River at Mason City, IA	1362	Major
East Fork Des Moines River near Algona, IA	2290	Major
Middle Raccoon River near Bayard, IA	971	Action
Wolf Creek near Dysart, IA	774	NA
Little Cedar River near Ionia, IA	793	Major
Volga River at Littleport, IA	901	Major
Yellow River at Necedah, WI	1272	Major
North Fork Maquoketa River near Fulton, IA	1308	Minor
Wapsipinicon River near Tripoli, IA	896	Major
Root River nr Pilot Mound, MN	1463	Major
Pecatonica River at Martintown, WI	2678	Major
Lemonweir River at New Lisbon, WI	1313	NA
Turkey River near Eldorado, IA	1660	Major
East Fork Des Moines River at Dakota City, IA	3388	Major
Pecatonica River at Freeport, IL	3437	Major
Cedar River at Charles City, IA	2730	Minor
Iowa River at Marshalltown, IA	3968	Major
Shell Rock River at Shell Rock, IA	4522	Major
Maquoketa River near Maquoketa, IA	4022	Major
Wapsipinicon River at Independence, IA	2714	Major
Turkey River above French Hollow Cr at Elkader, IA	2339	Minor
Cedar River at Waverly, IA	4007	Major
Cedar River at Janesville, IA	4302	Major

Table 2 Continued

Station	**Watershed area (km^2)**	**Max Stage**
Turkey River at Garber, IA	4002	Major
Maquoketa River nr Green Island, IA	4841	NA
Cedar River at Cedar Falls, IA	12261	Major
Iowa River near Belle Plaine, IA	6358	Minor
Cedar River at Waterloo, IA	13328	Major
Iowa River at Marengo, IA	7236	Major
Wapsipinicon River near Anamosa, IA	4079	Major
Cedar River at Vinton, IA	15644	Major
Black River nr Galesville, WI	5387	Major
Cedar River at Blairs Ferry Road at Palo, IA	16426	Major
Cedar River at Cedar Rapids, IA	16861	Major
Cedar River at Cedar Bluff, IA	18324	Major

Table 3: Cases used for this work.

03 May 2018 00z	18 June 2018 00z	24 August 2018 12z
04 May 2018 00z	18 June 2018 12z	26 August 2018 12z
12 May 2018 00z	20 June 2018 00z	27 August 2018 12z
14 May 2018 00z	20 June 2018 12z	28 August 2018 12z
23 May 2018 00z	21 June 2018 00z	02 September 2018 12z
14 June 2018 00z	24 June 2018 12z	03 September 2018 12z
16 June 2018 00z	26 June 2018 12z	04 September 2018 00z
16 June 2018 12z	30 June 2018 12z	04 September 2018 12z
17 June 2018 00z	01 July 2018 00z	05 September 2018 00z
17 June 2018 12z	19 August 2018 12z	

Table 4: Layout of a 2x2 contingency table.

		Observation	
		Yes	**No**
Forecast	**Yes**	Hit (*H*)	False Alarm (*FA*)
	No	Miss (*M*)	Correct Negative (*CN*)

Table 5: List of abbreviations and relevant equation numbers.

Metric	Abbreviation	Perfect Score	Equation Number(s)
FNE	Frequency of Non-Exceedance	1	1-2
RPS	Ranked Probability Score	0	3-5
POD	Probability of Detection	1	6
FAR	False Alarm Ratio	0	7
CSI	Critical Success Index	1	8
ETS	Equitable Threat Score	1	9-10
HSS	Heidke Skill Score	1	11
MAE	Mean Absolute Error	0	12

Table 6: Mean values for FNE and RPS across all events, for the four ensembles (when applicable) using equal weighting, weighting based on CI displacements, and weighting based on corrected CI displacements.

		Full Ensemble	Base HRRRE	Closest A	Closest B
Equal Weights	FNE	0.514	0.408	---	---
	RPS	0.423	0.454	---	---
CI Weighted	FNE	0.514	0.408	0.435	0.422
	RPS	0.429	0.465	0.441	0.447
CI Corrected	FNE	0.514	0.408	0.438	0.428
	RPS	0.428	0.469	0.441	0.448

Table 7: P-values gathered from a Two-Tailed Paired T-Test for the comparisons of the FNE and RPS values of all events for ensembles using equal weighting, weighting using CI displacements, and weighting using corrected CI displacements. The comparisons (when applicable) are the full ensemble to the base HRRRE (I-II), the full ensemble to the Closest A (I-III), the full ensemble to the Closest B (I-IV), the base HRRRE to the Closest A (II-III), the base HRRRE to Closest B (II-IV), and the Closest A to the Closest B (III-IV). Values have been bolded to show statistical significance at a 90% confidence interval.

		I – II	I – III	I – IV	II – III	II – IV	III – IV
Equal Weights	FNE	**< 0.0001**	---	---	---	---	---
	RPS	**0.0067**	---	---	---	---	---
CI Weighted	FNE	**< 0.0001**	**< 0.0001**	**< 0.0001**	**< 0.0001**	0.2383	0.1076
	RPS	**0.0022**	**0.0289**	**0.0010**	**0.0338**	**0.0945**	0.2259
CI Corrected	FNE	**< 0.0001**	**< 0.0001**	**< 0.0001**	**< 0.0001**	**0.0030**	**0.0203**
	RPS	**0.0001**	**0.0069**	**< 0.0001**	**0.0031**	**0.0362**	**0.0933**

Table 8: MAE for all ensembles and weighting schemes as a guide to quantifying reliability. Lower values show that the distribution fell closer to theoretical perfect, and are thus better values.

	Full Ensemble	Base HRRRE	Closest A	Closest B
Equal Weights	0.150	0.223	---	---
CI Weighted	0.174	0.234	0.206	0.198
Corr. CI. Weighted	0.174	0.233	0.216	0.203

Table 9: POD, FAR, CSI, ETS, and HSS for each POE threshold (%) using the full ensemble (I) and base HRRRE (II) ensemble forecasts with equal weighting. The peak values for each skill metric and each ensemble have been bolded.

Eq. Weights		> 0	> 10	> 20	> 30	> 40	> 50	> 60	> 70	> 80	> 90
POD	I	0.888	0.771	0.659	0.659	0.603	0.575	0.559	0.503	0.475	0.436
	II	0.816	0.816	0.760	0.709	0.682	0.620	0.592	0.553	0.514	0.486
FAR	I	0.762	0.644	0.567	0.494	0.383	0.299	0.248	0.159	0.115	0.093
	II	0.650	0.650	0.568	0.517	0.453	0.373	0.312	0.244	0.185	0.130
CSI	I	0.231	0.322	0.368	0.401	0.439	0.462	**0.472**	0.459	0.447	0.417
	II	0.324	0.324	0.380	0.403	0.436	0.453	0.467	**0.469**	0.460	0.453
ETS	I	0.107	0.220	0.280	0.325	0.375	0.406	**0.419**	0.413	0.404	0.376
	II	0.221	0.221	0.290	0.321	0.363	0.390	0.409	**0.417**	0.412	0.409
HSS	I	0.193	0.361	0.438	0.490	0.546	0.577	**0.591**	0.585	0.576	0.546
	II	0.362	0.362	0.450	0.487	0.533	0.561	0.581	**0.589**	0.584	0.580

Table 10: As in Table 9 but for the uncorrected CI displacement error scheme, now including the Closest A ensemble (III) and Closest B ensemble (IV). Peak values for each skill metric and each ensemble have been bolded.

CI Weighted		> 0	> 10	> 20	> 30	> 40	> 50	> 60	> 70	> 80	> 90
POD	I	0.888	0.782	0.704	0.665	0.631	0.598	0.564	0.536	0.469	0.447
	II	0.816	0.788	0.749	0.704	0.670	0.626	0.592	0.581	0.531	0.503
	III	0.799	0.799	0.732	0.704	0.687	0.642	0.587	0.536	0.497	0.453
	IV	0.827	0.827	0.743	0.704	0.682	0.642	0.598	0.536	0.497	0.430
FAR	I	0.762	0.634	0.570	0.498	0.429	0.323	0.279	0.186	0.125	0.091
	II	0.650	0.604	0.553	0.504	0.447	0.385	0.350	0.302	0.234	0.159
	III	0.690	0.690	0.589	0.519	0.451	0.361	0.295	0.232	0.144	0.100
	IV	0.670	0.670	0.584	0.519	0.440	0.382	0.291	0.226	0.152	0.083
CSI	I	0.231	0.332	0.364	0.401	0.428	0.465	0.463	**0.478**	0.440	0.428
	II	0.324	0.358	0.388	0.410	0.435	0.450	0.449	**0.464**	0.457	0.459
	III	0.287	0.287	0.357	0.400	0.439	**0.471**	**0.471**	0.462	0.459	0.431
	IV	0.309	0.309	0.363	0.400	0.444	0.460	**0.480**	0.464	0.456	0.414
ETS	I	0.107	0.232	0.276	0.323	0.359	0.407	0.409	**0.430**	0.396	0.386
	II	0.221	0.262	0.301	0.331	0.363	0.385	0.389	0.408	0.406	**0.413**
	III	0.178	0.178	0.265	0.319	0.367	0.409	**0.415**	0.411	0.414	0.389
	IV	0.202	0.202	0.272	0.319	0.372	0.395	**0.424**	0.413	0.411	0.373
HSS	I	0.193	0.376	0.433	0.488	0.528	0.578	0.580	**0.601**	0.568	0.557
	II	0.362	0.415	0.463	0.497	0.533	0.556	0.560	0.579	0.577	**0.585**
	III	0.302	0.302	0.419	0.483	0.536	0.580	**0.586**	0.582	0.585	0.560
	IV	0.336	0.336	0.428	0.483	0.542	0.567	**0.595**	0.585	0.582	0.544

Table 11: As in Table 10 but for the corrected CI displacement error scheme. Peak values for each skill metric and each ensemble have been bolded.

CI Corr.		> 0	> 10	> 20	> 30	> 40	> 50	> 60	> 70	> 80	> 90
POD	I	0.888	0.788	0.721	0.682	0.631	0.603	0.564	0.531	0.480	0.447
	II	0.816	0.760	0.749	0.704	0.670	0.631	0.581	0.570	0.525	0.503
	III	0.810	0.810	0.749	0.704	0.654	0.620	0.570	0.547	0.492	0.469
	IV	0.816	0.816	0.749	0.698	0.682	0.620	0.598	0.547	0.503	0.441
FAR	I	0.762	0.645	0.560	0.487	0.405	0.303	0.263	0.208	0.122	0.091
	II	0.650	0.617	0.564	0.504	0.450	0.386	0.325	0.227	0.223	0.189
	III	0.691	0.691	0.585	0.508	0.443	0.373	0.306	0.203	0.162	0.097
	IV	0.676	0.676	0.595	0.532	0.443	0.373	0.282	0.240	0.174	0.092
CSI	I	0.231	0.324	0.376	0.414	0.441	**0.478**	0.470	0.466	0.450	0.428
	II	0.324	0.342	0.381	0.410	0.433	0.452	0.454	**0.488**	0.456	0.450
	III	0.288	0.288	0.364	0.408	0.430	0.453	0.455	**0.480**	0.449	0.447
	IV	0.302	0.302	0.356	0.389	0.442	0.453	**0.484**	0.467	0.455	0.422
ETS	I	0.107	0.222	0.289	0.337	0.375	**0.421**	0.416	0.416	0.407	0.386
	II	0.221	0.245	0.292	0.331	0.361	0.387	0.396	**0.437**	0.406	0.402
	III	0.178	0.178	0.272	0.328	0.359	0.390	0.399	**0.431**	0.403	0.405
	IV	0.194	0.194	0.263	0.307	0.370	0.390	**0.429**	0.415	0.408	0.381
HSS	I	0.193	0.364	0.448	0.504	0.545	**0.592**	0.588	0.588	0.578	0.557
	II	0.362	0.394	0.452	0.497	0.531	0.558	0.567	**0.608**	0.578	0.574
	III	0.303	0.303	0.428	0.494	0.528	0.561	0.570	**0.603**	0.575	0.576
	IV	0.325	0.325	0.417	0.469	0.540	0.561	**0.600**	0.587	0.579	0.552

Table 12: As in Table 7 but for POD, FAR, CSI, ETS, and HSS. Values have been bolded to show statistical significance at a 90% confidence interval.

		I - II	I - III	I - IV	II - III	II - IV	III - IV
Equal Weights	POD	**0.0176**	---	---	---	---	---
	FAR	0.1186	---	---	---	---	---
	CSI	0.1474	---	---	---	---	---
	ETS	0.3144	---	---	---	---	---
	HSS	0.3128	---	---	---	---	---
CI Weighted	POD	**0.0510**	0.2687	0.1089	0.1476	0.5080	0.3238
	FAR	0.2234	0.1405	0.3740	0.5423	0.3326	0.2155
	CSI	0.1368	0.6833	0.4609	0.1854	0.2543	0.3955
	ETS	0.2539	0.8615	0.6043	0.2387	0.3134	0.3557
	HSS	0.2375	0.8223	0.5622	0.1981	0.2522	0.2792
CI Corrected	POD	0.1921	0.3899	0.2366	0.2960	0.6934	0.4054
	FAR	0.1958	0.1172	0.1110	0.6287	0.7432	0.6789
	CSI	0.3055	0.8619	0.9481	**0.0524**	0.1057	0.8713
	ETS	0.4466	0.7535	0.8212	**0.0995**	0.1278	0.9009
	HSS	0.3903	0.9013	0.9853	0.1068	0.1046	0.8148

Table 13: Normalized percentage of ensemble members' weights distributed by the ensemble members' timing of peak discharge. The timing categories have been broken up into greater than 6 hours early, 4-6 hours early, 1-3 hours early, timing hits (hour 0), 1-3 hours early, timing hits (hour 0), 1-3 hours late, 4-6 hours late, and greater than 6 hours late. All values are adjusted to only include members that surpassed action stage before peaking for the full ensemble (I), base HRRRE (II), Closest A (III) and Closest B (IV). All weighting schemes have been included, as well as mean timing for all ensembles and weighting schemes.

		< 6 Early	4-6 Early	1-3 Early	Hit	1-3 Late	4-6 Late	> 6 Late
Equal Weights	I	46.92	5.14	4.99	1.01	3.21	7.05	31.68
	II	46.97	5.12	4.83	1.05	3.43	6.57	32.04
CI Weighted	I	47.12	4.94	5.08	1.27	3.10	7.01	31.48
	II	46.08	5.52	4.75	1.19	3.41	7.14	31.91
	III	44.84	5.80	5.26	1.66	3.30	7.41	31.72
	IV	46.94	5.46	5.06	1.56	2.64	6.77	31.58
CI Corrected	I	46.66	5.28	4.77	0.99	3.39	6.82	32.10
	II	46.09	5.62	4.16	1.13	3.38	6.86	32.77
	III	46.48	5.47	4.89	1.44	3.85	7.36	30.51
	IV	44.83	6.12	4.68	0.97	3.65	7.13	32.62
Mean		46.29	5.45	4.85	1.23	3.33	7.01	31.84

3.10 Figures

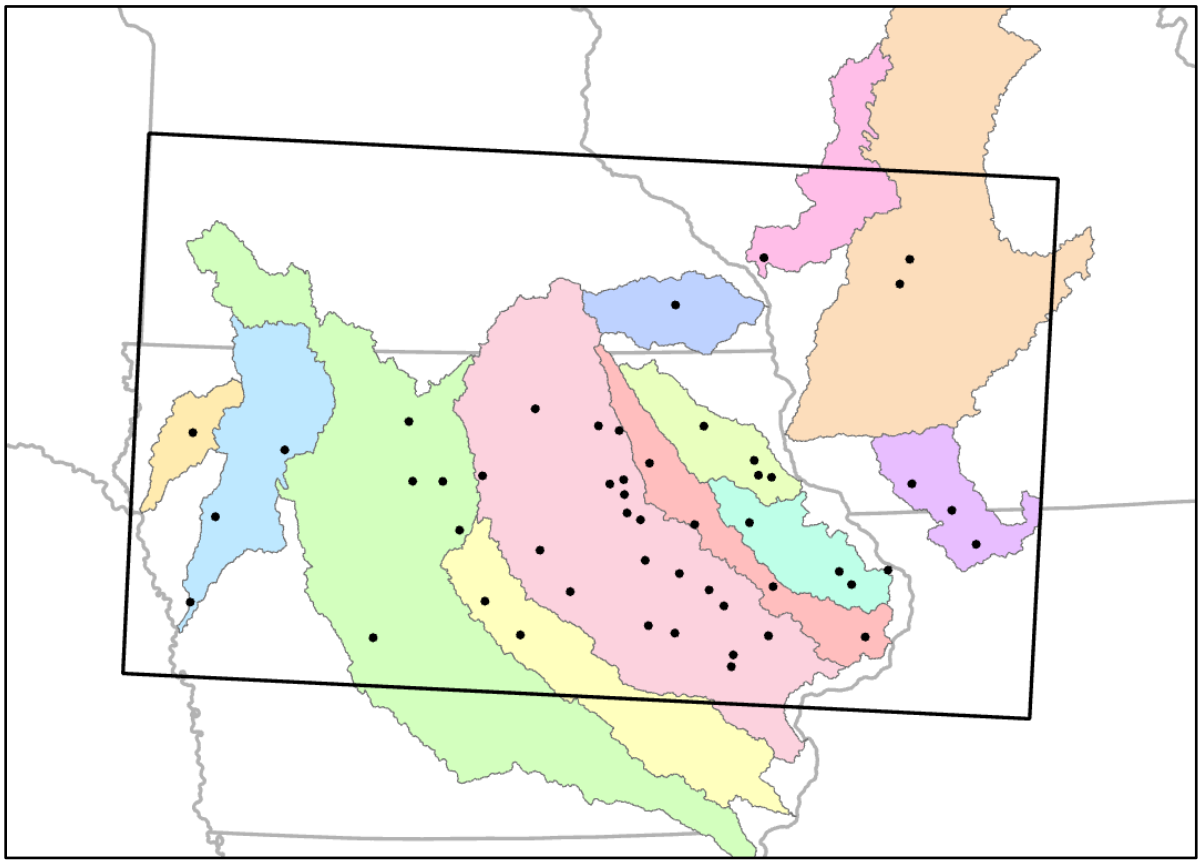


Figure 1: NWM domain obtained from NCAR (bolded rectangle). Gauges that fit the requirements for Pbias and NSE for the 2018 summer season are also shown, with their major watersheds shown in colored polygons.

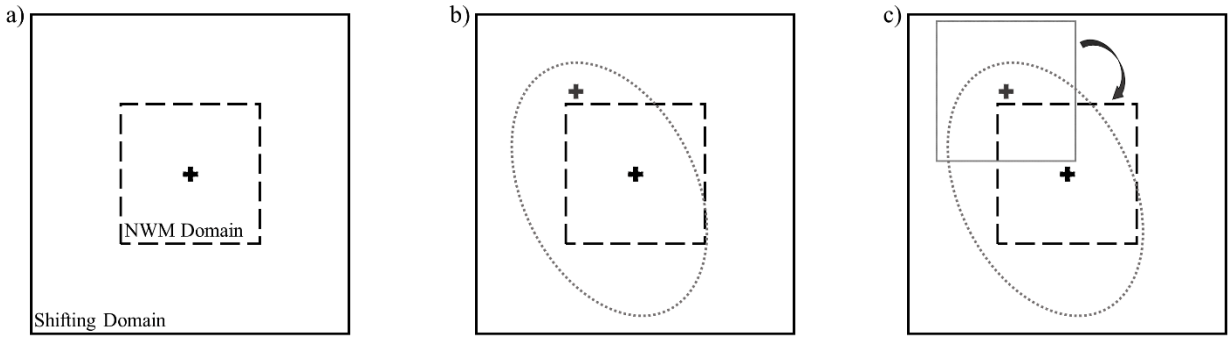
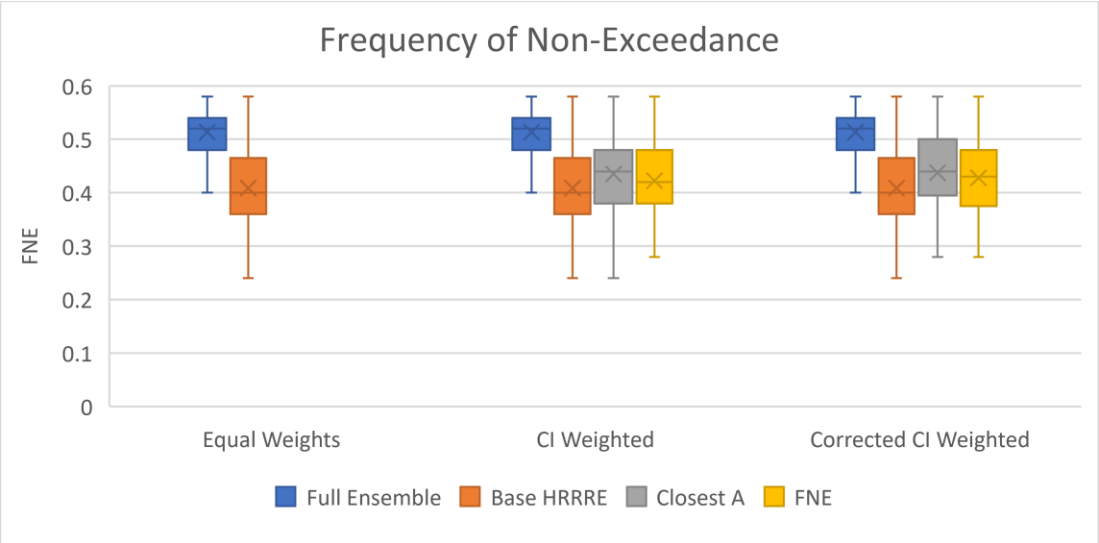


Figure 2: A visual example of the methodology for shifting QPFs within the WRF-Hydro framework. a) The solid black square shows the shifting domain. The long dashed black square is the NWM domain. The black cross is the center point of the NWM domain. b) The gray dashed oval graphically represents the climatological displacement distributions found by Kiel et. al. (2021) The snRNG picks 54 latitude and longitude points from inside the area of the dashed oval. Those latitude and longitude points then become the centers of a shifted domain (gray cross). c) Data that matches the dimensions of the NWM cutout, centered at the gray cross, are then moved to the center of the NWM cutout (long dashed black square). This is done to deceive the model into using the precipitation from the shifted domain (gray solid square) while having the coordinates and familiar grid of the NWM domain (long dashed black square).

a)



b)

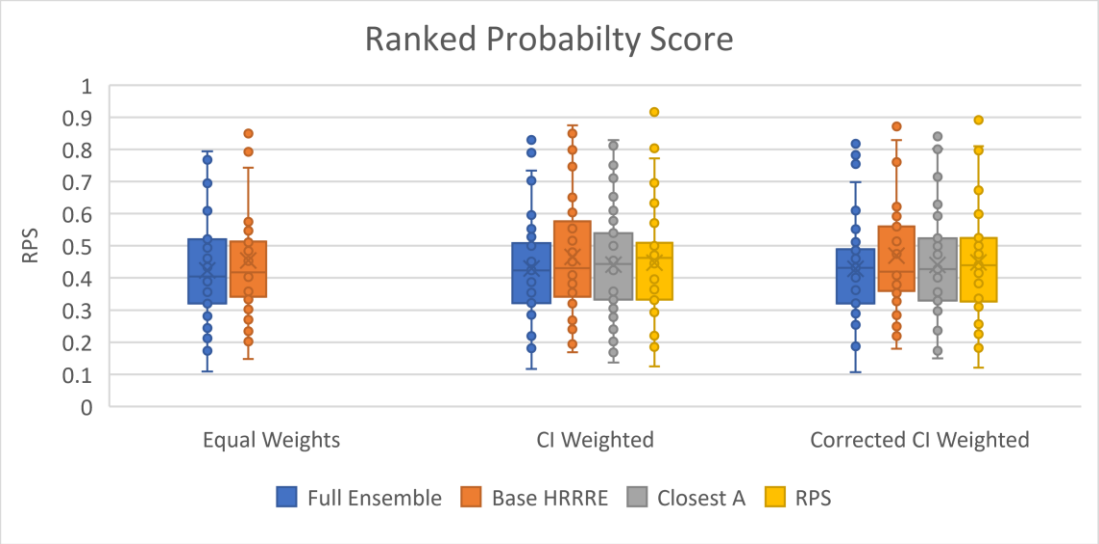
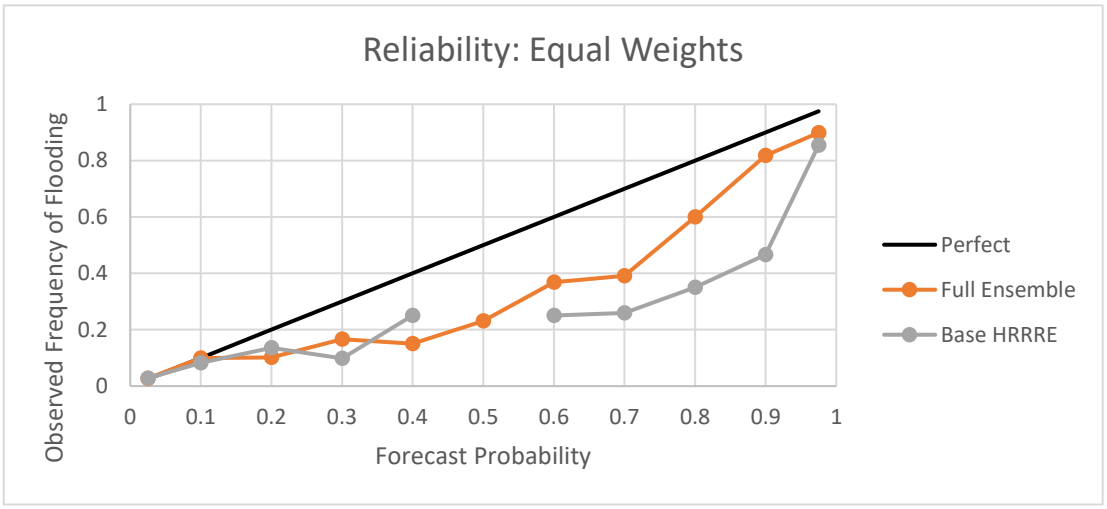
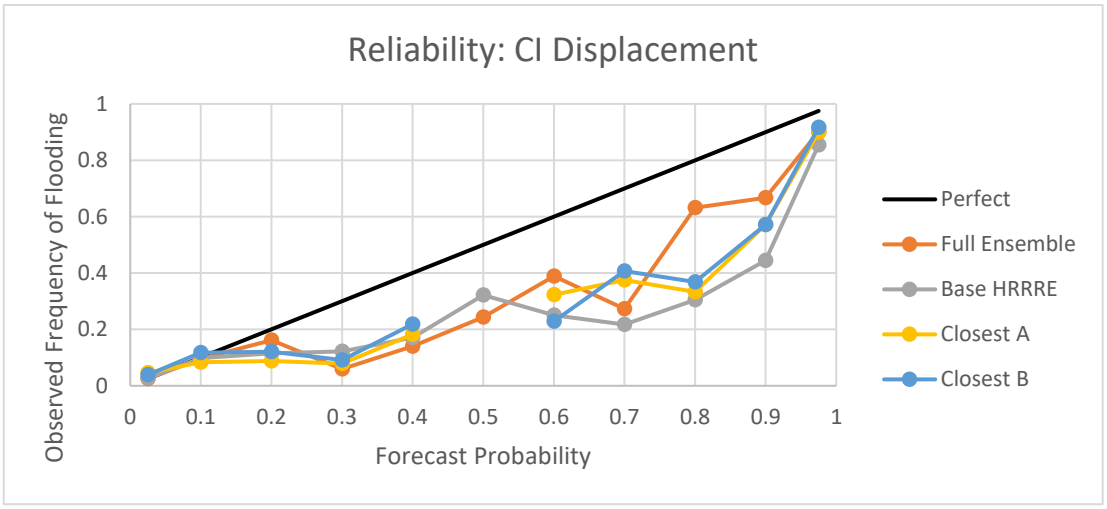


Figure 3: Box and whisker plots for the a) frequency of non-exceedance (FNE) and b) ranked probability score (RPS) for all ensembles with the three weighting schemes: equal weighting, weighting based on the uncorrected CI displacements, and weighting based on corrected CI displacements.

a)



b)



c)

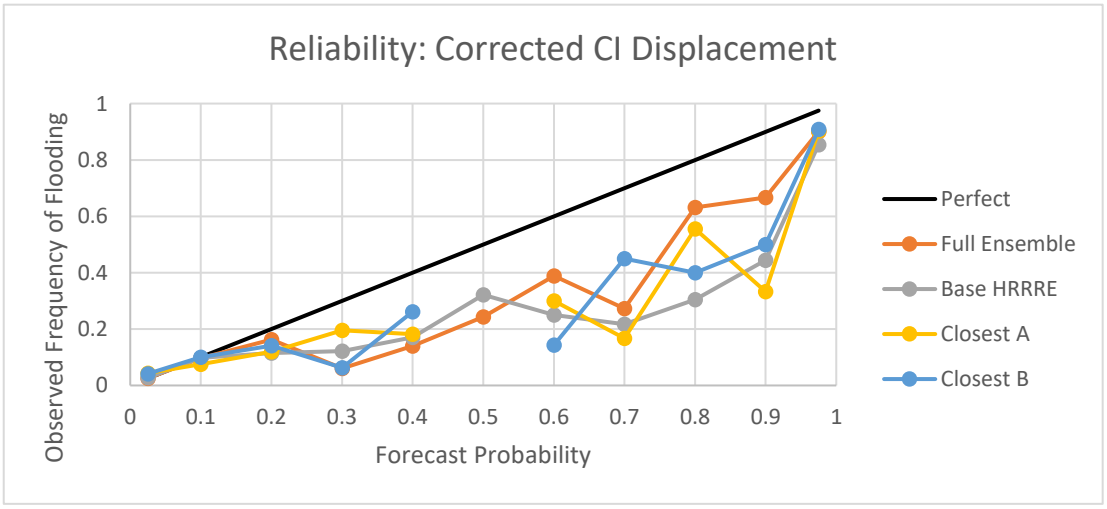


Figure 4: Reliability curves for four ensembles (colors explained in legend on right) using a) equal weighting, b) weights generated by accounting for CI displacements, and c) the weights created using the Kiel et al. (2021) correction for the CI displacements. A line of theoretically perfect reliability has been added (solid black) to provide a reference.

CHAPTER 4. GENERAL CONCLUSIONS

4.1 Conclusions

This thesis tested the use of climatological spatial displacements of predicted precipitation to guide shifting of QPF for improved ensemble streamflow forecasts. The full 63-member ensemble, made of shifted and non-shifted members, showed promise by having better deterministic skill measures for predictions of peak flow when using equal weighting and uncorrected CI displacement weights. However, when using the corrected CI displacement weights, the base HRRRE produced the best values for CSI, ETS, and HSS of any ensemble. Additionally, when averaging across all POE thresholds, the base HRRRE consistently produced the best skill values. Moreover, due to lack of statistical significance at a 90% confidence interval for these measures, none of the ensembles can be definitively identified as having more skill in deterministic forecasting over the others. For our probabilistic measures of FNE and RPS, the full ensemble had statistically better scores than the other three ensembles, for all weighting schemes. Due to this fact, it appears that the use of a climatological spatial displacement dataset, like the one found in Kiel et al. (2021), as a driver for spatial shifting of HRRRE QPF produced allowed for the improvement of probabilistic forecasts, with relatively equal deterministic forecasting ability, as compared to the non-shifted original HRRRE members.

For the probabilistic measures, the two Closest A and B ensembles generally performed better than the base HRRRE but fell short of the full ensemble. Similarly, their average deterministic skills scores were greater than the full ensemble while also being worse than the base HRRRE. For the deterministic measures, there were no instances of statistical significance to support them being better than the full ensemble, whereas the full ensemble was statistically superior to the Closest A and B ensembles for FNE and RPS. There was also an absence of

statistical differences between the two Closest A and B methodologies. This led to the conclusion that these methods require refinement improve upon the base HRRRE and the full ensemble.

There were only a few instances where statistically significant differences occurred between weighting schemes. One of those occasions was for the FAR of the full ensemble. When using either of the two CI displacement weighting schemes the FAR got significantly worse using a 90% confidence interval. However, when the corrected weights were used, the full ensemble saw its best mean POD. There was no significance in the CSI, ETS, and HSS skill values for the 90% confidence interval. For all of the ensembles studied, there were no significant improvement or deterioration in their deterministic skill metrics when switching between weighting schemes.

When evaluating reliability, the two CI displacement weighting schemes had the lowest MAE scores. However, none of the changes in reliability between schemes were significant at a 90% confidence interval. The FNE for the Closest B increased when using the corrected CI displacement errors. It was a significant increase over the FNE for the uncorrected CI displacement errors. When looking at RPS, the base HRRRE's poor output when using the two CI displacement weighting schemes were statistically significant. So too was the worsening of RPS for the full ensemble when using the two CI displacement weighting schemes. These data suggest that the weighting schemes did not improve the skill of the ensembles, though there were limited instances of statistical significance.

For the comparison between "large" and "small" basins, the larger basins had significantly higher deterministic skill metric scores at all POE thresholds at a 90% confidence interval. For probabilistic measures, larger basins produced better RPS values. The only metric

that showed smaller basins having better scores over larger basins was FNE. Both findings for RPS and FNE were statistically significant.

The analysis for the timing of peak discharge showed that all four ensembles performed poorly. On average, 78% of the ensemble member weights were over 6 hours early or 6 hours late. Interestingly, the shifting of QPF appears to not have influenced the timing of peak discharge, as there were no statistically significant comparisons between ensembles. Given the data found here, it cannot be determined where these issues in prediction of flood peak timing originate in the modelling process.

Overall, this methodology shows promise in its ability to expand a streamflow ensemble quickly and efficiently, while also helping to account for the uncertainty in flood forecasting created by climatological displacement errors present in numerical weather models. This method has shown itself to be most beneficial for flood prediction in larger basins.

4.2 Future Work

The results in Chapter 3 showed the depth and span of work done to improve ensemble streamflow prediction using spatially shifted QPF, however there are several avenues for additional research. The weighting methods based on CI displacements of QPF were unable to produce consistent improvement and, in many cases, worsened skill scores. A sensitivity analysis could be used to identify an empirical exponent or coefficient to increase the effectiveness of the CI displacement weighting schemes. In the third weighting scheme, the correction devised by Kiel et al. (2021) to reduce the sometimes-unrealistically large CI displacements was used to better account for the accumulated QPF displacements seen through forecast hour 18 of the HRRRE model runs. When shifted members were included in the ensembles, the correction was not effective. Of course, it is important to remember that the correction was designed to improve the base HRRRE members, and it only accounts for the first 18 hours of accumulated

displacement. Although the rain events used in this study had most of the precipitation falling within the first 18 hours of the forecast period, our QPFs were still allowed to run out to the end of the 36-hour HRRRE forecast. The reductions in CI displacement may be less relevant as the displacement accumulation continues through forecast hour 36. More research needs to be completed to investigate the Kiel et al. (2021) correction for CI displacements and its use for longer forecast periods, as well as how it can be better used to refine shifting methods that are based on climatological spatial displacements.

The Closest A and B ensembles showed deterministic performance on par with the full ensemble and the base HRRRE, while having slightly worse scores that the full ensemble when considering probabilistic metrics. They likely require more fine-tuning to increase their effectiveness as forecasting tools. This may involve an empirical exponent or coefficient to modify the CI displacement errors like described previously. Another option that was not explored in this work that may improve the Closest A methodology could take advantage of the average streamflow values from the parent HRRRE member groupings. Seeing the results of this work, if the Closest A and B ensembles are expanded to include more members, one or both of these two ensembles may produce similar deterministic skill values to the base HRRRE, while still having some of the probabilistic forecasting ability of the larger, full ensemble.

Additional work is required to identify which portion of the modelling process is contributing the most to temporal errors in simulated flood waves. Only then could the shifting methods be adjusted to account for temporal and spatial displacement errors. The timing of CI was not examined in this thesis but may provide insight into the origin of these issues. This suggested investigation could be aided using object-based evaluations of HRRRE QPF and

watershed orientation. It is likely that small errors in MCS orientation could cause issues in the prediction of flood peak timing, as well as the poor ensemble performance for small basins.

APPENDIX. SUMPLIMENTARY INFORMATION

$$NSE = 1 - \left(\frac{\sum_{j=1}^{m} \left(f_j - q_j \right)^2}{\sum_{j=1}^{m} (q_j - Q)} \right) \quad m = 1, \dots, J \tag{a}$$

$$Pbias = \left(\frac{\sum_{j=1}^{m} \left(f_j - q_j \right)}{\sum_{j=1}^{m} (q_j)} \right) * 100\% \quad m = 1, \dots, J \tag{b}$$

Figure A.1: The equations above were used to identify gauges with good performance when using reanalysis NLDAS-2 data as input to the WRF-Hydro in the NWM configuration, where *fj* is the forecast at time *j*, q_j is the measured discharge at time *j*, and *Q* is the mean observed discharge. These metrics were calculated for all stream gauges. a) is the Nash-Sutcliffe efficiency (Nash and Sutcliffe 1970) b) is the percent bias.

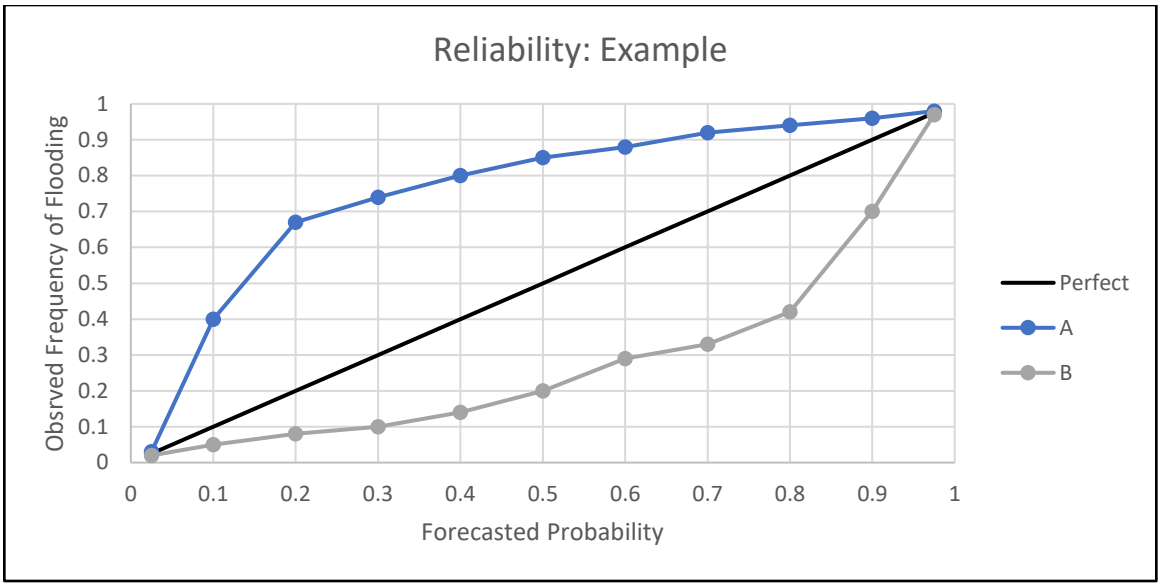
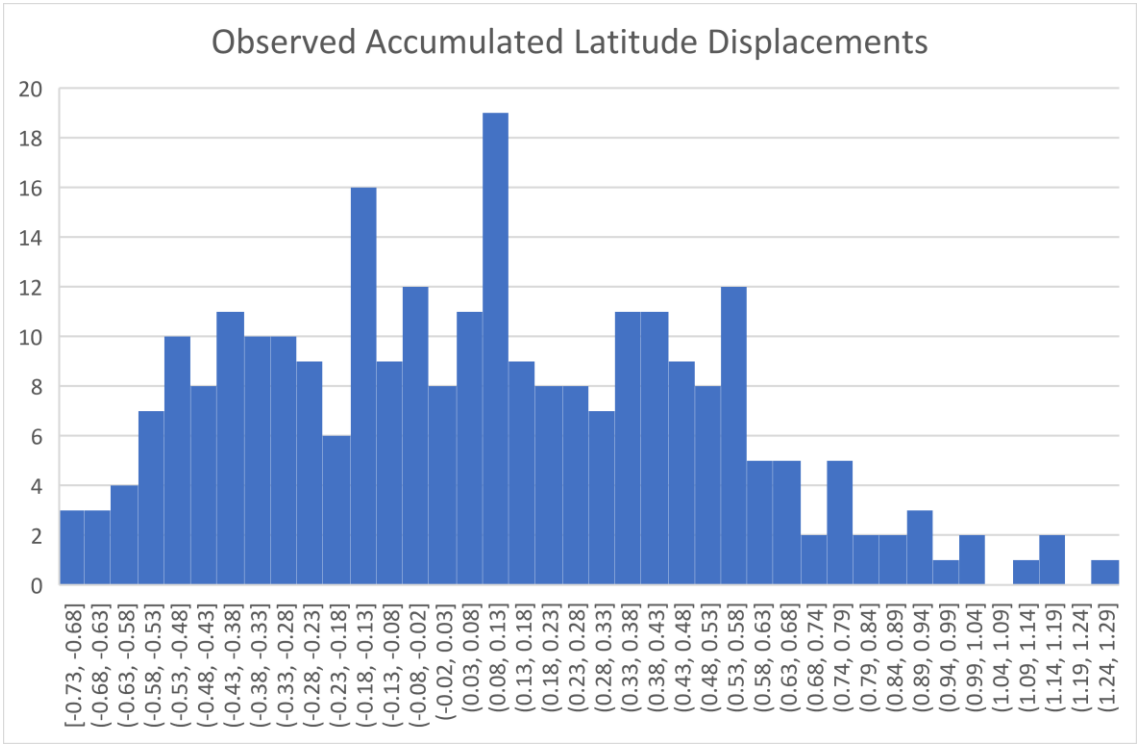


Figure A.2: An example of a set of reliability curves. The theoretical perfect forecasting ability is shown as the solid black line. Curve A shows a model "under forecasting" flood events, whereas curve B shows "over forecasting."

a)



b)

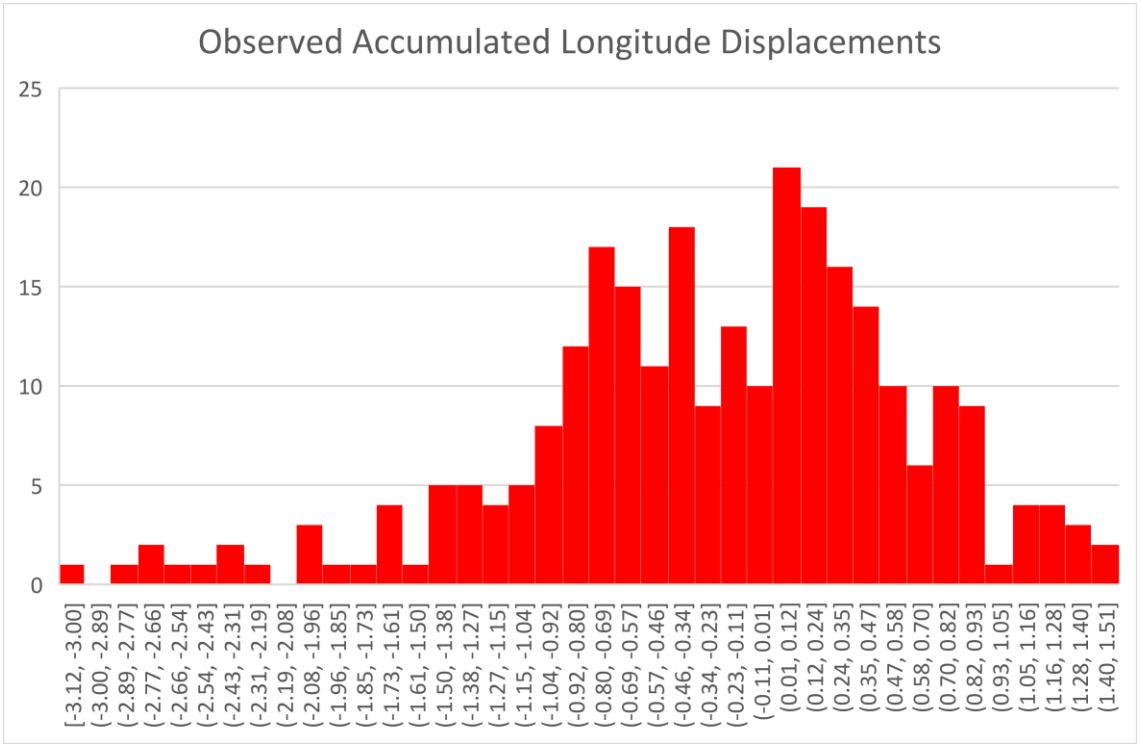


Figure A.3 a) Accumulated 0-18 hour forecast displacement between QPF centroids and observed precipitation centroids. In values of degrees latitude. b) same as a) but for longitude displacements. Values are in degrees longitude.

Table A.1: This table is adapted from Table 7 found in Kiel et al. (2021). It shows the suggested corrections for CI displacements in the North-South direction (N-S) and East-West direction for CI displacements in all four quadrants. All mean reduction values are in kilometers. A two-tailed paired T-test provided the p-values shown here. Values are bolded to show significance at a 95% confidence interval.

Quadrant	N-S Mean Reduction	N-S P-values	E-W Mean Reduction	E-W P-values
NW:	**13.32**	**0.008**	5.04	0.495
NE:	**10.58**	**0.006**	8.84	0.126
SW:	1.39	0.19	**7.12**	**0.035**
SE:	**5.65**	**0.035**	-1.22	0.446