Assessing and enhancing the West Gulf River Forecast Center's Hydrologic Ensemble Forecast Service for Forecast Informed Reservoir Operation Pilot Projects in Texas

Yu Zhang, Ph. D
Hue Nguyen, Ph. D
Robel Geressu, Ph. D
Yanjun Gan, Ph. D
University of Texas at Arlington

Texas Water Development Board Contract 2300012685

Final Report: Assessing and enhancing the West Gulf River Forecast Center's Hydrologic Ensemble Forecast Service for Forecast Informed Reservoir Operation Pilot Projects in Texas

TWDB Contract Number 2300012685

This page is intentionally blank.

List of Acronyms

AORC	Analysis of Record for Calibration
BS	Brier Score
BSS	Brier Skill Score
CBPR	Conditional Bias Penalizing Regression
CDF	Cumulative Density Function
CFSv2	Climate Forecast System version 2
CRPS	Continuously Ranked Probability Score
CRPSS	Continuously Ranked Probability Skill Score
CSGD	Censored-Shifted Gamma Distribution
ECMWF	European Center for Medium-range Weather Forecast
EnsPost	Ensemble Postprocessor
ESP	Ensemble Streamflow Prediction
EVS	Ensemble Verification System
FIRO	Forecast-Informed Reservoir Operation
GEFS	Global Ensemble Forecast System
HEFS	Hydrologic Ensemble Forecast Service
HRRR	High-resolution rapid refresh
KGE	King-Gupta Efficiency
LRS	Little River System
MAP	Mean Areal Precipitation
MEFP	Meteorological Ensemble Forcing Processor
MMGD	Mixed Meta-Gaussian Distribution
NOAA	National Oceanic and Atmospheric Administration
NIDIS	National Integrated Drought Information System
NSE	Nash-Sutcliffe Efficiency
NWP	Numerical Weather Prediction
NWS	National Weather Service
OWP	Office of Water Prediction
PSL	Physical Science Lab
ROC	Receiver Operating Characteristic
S2S	Subseasonal to Seasonal
SAC-SMA	Sacramento Soil Moisture Accounting
TWDB	Texas Water Development Board
USACE	United States Army Corps of Engineers
USACE-SWF	United States Army Corps of Engineers - Fort Worth District
USBR	US Bureau of Reclamation
USGS	US Geological Survey
UTA	University of Texas at Arlington
WGRFC	West Gulf River Forecast Center

Table of Contents

List of Tables

List of Figures

Executive Summary

The first Forecast-informed Reservoir Operations (FIRO) pilot project in Texas (referred to hereafter as "FIRO Pilot") was launched in 2022. It is led by Hydrometeorology Research Group at the University of Texas at Arlington (UTA). The domain of the pilot is situated in Central Texas within the Brazos River Basin, encompassing the lower portion of the Little River System (LRS) that houses a system of multiuse reservoirs operated by the U.S. Army Corps of Engineers (USACE).

The first phase of the Texas FIRO Pilot was supported by a grant from the US Bureau of Reclamation (USBR)'s WaterSmart Program. In 2023, TWDB provided a supplemental grant to the project team with the following objectives:

- Updating the configurations of the Hydrologic Ensemble Forecast Service (HEFS) and reservoir model from National Weather Service (NWS) West Gulf River Forecast Center (WGRFC) to be consistent with the latter's real-time river operations and more accurately represent reservoir operations.
- Expanding the hindcast and validation efforts to inform WGRFC on potential improvements to the configurations to yield more skillful forecasts.
- Providing forecast data and support to TWDB's Lake Conroe initiative.

The project comprises the following five tasks:

1. Review and revise the hindcast configuration of WGRFC HEFS for the Little River System.
2. Produce hindcasts for extended lead times using the revised versions of WGRFC HEFS and perform validation.
3. Perform validation on Climate Forecast System-version 2 (CFSv2) and Global Ensemble Forecast System-version 12 (GEFSv12) subseasonal precipitation forecasts.
4. Deliver postprocessed ensemble precipitation and streamflow forecasts upstream Lake Conroe.
5. Report findings to stakeholders through regular meetings.

Key outcomes from the project are summarized as follows

1. Developed of four updated HEFS stand-alone configurations for the Little River System to produce three suites of hindcasts at 3-hour intervals. These configurations incorporate updated basin boundaries and Sacramento Soil Moisture Accounting (SAC-SMA) parameter values from the most recent round of calibration completed in 2022 These configurations differ in the forcing input:
    a. Configuration 1: with resampled climatology as the sole source of precipitation forcing.
    b. Configuration 2: with GEFSv12 medium-range precipitation forecasts for day 1-14 and resampled climatology for day 15–270.
    c. Configuration 3: with GEFSv12 subseasonal to seasonal (S2S) precipitation forecasts for day 1–35 and resampled climatology for day 36–270.

d. Configuration 4: with CFSv2 precipitation forecasts for day 15-270.

2. Generation of validation statistics for forecast points collocated with USGS gauging stations, and at inflow points to four reservoirs. Notable findings include the following:
    a. As judged by summary statistics, ensemble streamflow hindcasts driven by GEFSv12 medium-range forecasts are more skillful than those driven by resampled climatology for lead time out to day 20. When aggregated onto weekly scales, they are more skillful than climatology for lead times of 1-3 months. The streamflow hindcasts tend to be skillful at longer lead times than precipitation forecasts due to the dampening effects of runoff process.
    b. There is a north-south gradient in the skills of ensemble streamflow forecasts: the skills tend to be higher for forecast points in the north and decline southward. Potential sources of this gradient include larger drainage area and slower runoff response in the northern parts of LRS, and flashier response and prevalence of convection with low predictability in the southern parts.
    c. Forecast skills are dependent on flow magnitude. Most of the skills reside in forecasting moderate and moderate-high flows, whereas for low flows the former consistently underperforms resampled climatology-driven hindcasts.
    d. High flows during major floods are systematically under-forecasted. The under-forecast can be attributed to a) under-forecasts in precipitation amounts, and b) issues in the water balance model in reproducing the runoff volume.
    e. The issue of under-forecast tends to be more severe when the newer set of SAC-SMA parameter values were ingested.
    f. Ensemble streamflow forecasts driven by GEFSv12 S2S forecasts and resample climatology exhibit higher skills than those driven solely by resampled climatology for major flooding events.
    g. Ensemble streamflow forecasts driven by using CFSv2 are not as skillful as those driven by resampled climatology across lead time, likely due to a lack of skills in CFSv2 precipitation forecasts and inadequacy of the postprocessing technique to calibrate these forecasts.

3. Development of HEFS ensemble precipitation and flow forecasts for the San Jacinto River Basin upstream Lake Conroe for June-September of 2005 2006, 2017, and 2018. This task was solely to provide the forecasts to the TWDB, and no further analysis was undertaken.

On the basis of the findings, the project team makes the following recommendations to both the NWS and the TWDB to improve HEFS to facilitate the use of its ensemble streamflow forecasts:

1. Improve forecast skills for anomalously large precipitation events in the medium range (days 1 – 12), possibly through.
    - introducing alternative, advanced postprocessing algorithms to HEFS to alleviate the negative bias and under-spread in raw precipitation forecasts. Candidate algorithms include Conditional Bias Penalizing Regression, Censored-Shifted Gamma Distribution and a later machine-learning variant, and alternatives to Schaake Shuffle.

    - exploring alternative precipitation forecasts as input to HEFS, such as European Center High-resolution forecasts and operational convection-allowing model forecasts.
2. Improve the calibration of NWS hydrologic and routing models to better capture the magnitude of inflow during flood events.
    - Recalibrate SAC-SMA for forecast points in the pilot domain by including specific metrics for large events.
    - Quality control reservoir inflow estimates from USACE.
    - Determine optimal precipitation products for calibration and verification, including an assessment of the value of the raw Analysis of Record for Calibration (AORC) and the bias-corrected AORC, which was developed by UTA.
3. Investigate alternative forecast products as forcing to HEFS beyond the medium range. These include subseasonal forecasts from GEFS, and North American Multimodel Ensemble.
4. Partner with reservoir operators to add and refine metrics for forecast evaluation and improvements to facilitate consistent use of forecasts in operational decisions. It is recommended that the project team partner with reservoir operators to identify scenarios where the impacts from potential failures of forecasts can be gauged and accounted for. The National FIRO Program now uses Critical Success Index and Dry Forecast Failure ratio as metrics in its screening of reservoirs, and their inclusion in future analysis is recommended.

1. Background

In 2019, UT Arlington (UTA) collaborated with the Texas Water Development Board (TWDB) and the National Integrated Drought Information System (NIDIS) to convene the first Texas-Oklahoma Forecast Informed Reservoir Operation (FIRO) Workshop. The report of the workshop called for establishing FIRO pilot sites in the state of Texas, whereby existing and emerging National Weather Service (NWS) forecast products can be tested in a quasi-operational setting to allow for the identification and adoption of mature, robust products in reservoir operation (TWDB, 2020). In 2020, a multi-agency consortium was formed to facilitate the introduction of FIRO practice in the state of Texas, comprising the UTA, US Army Corps of Engineers – Fort Worth District (USACE-SWF), NWS, and the TWDB. In 2021, USACE-SWF helped identify a potential pilot site in central Texas in the Little River System, a tributary to the Brazos River (Fig 1-1).

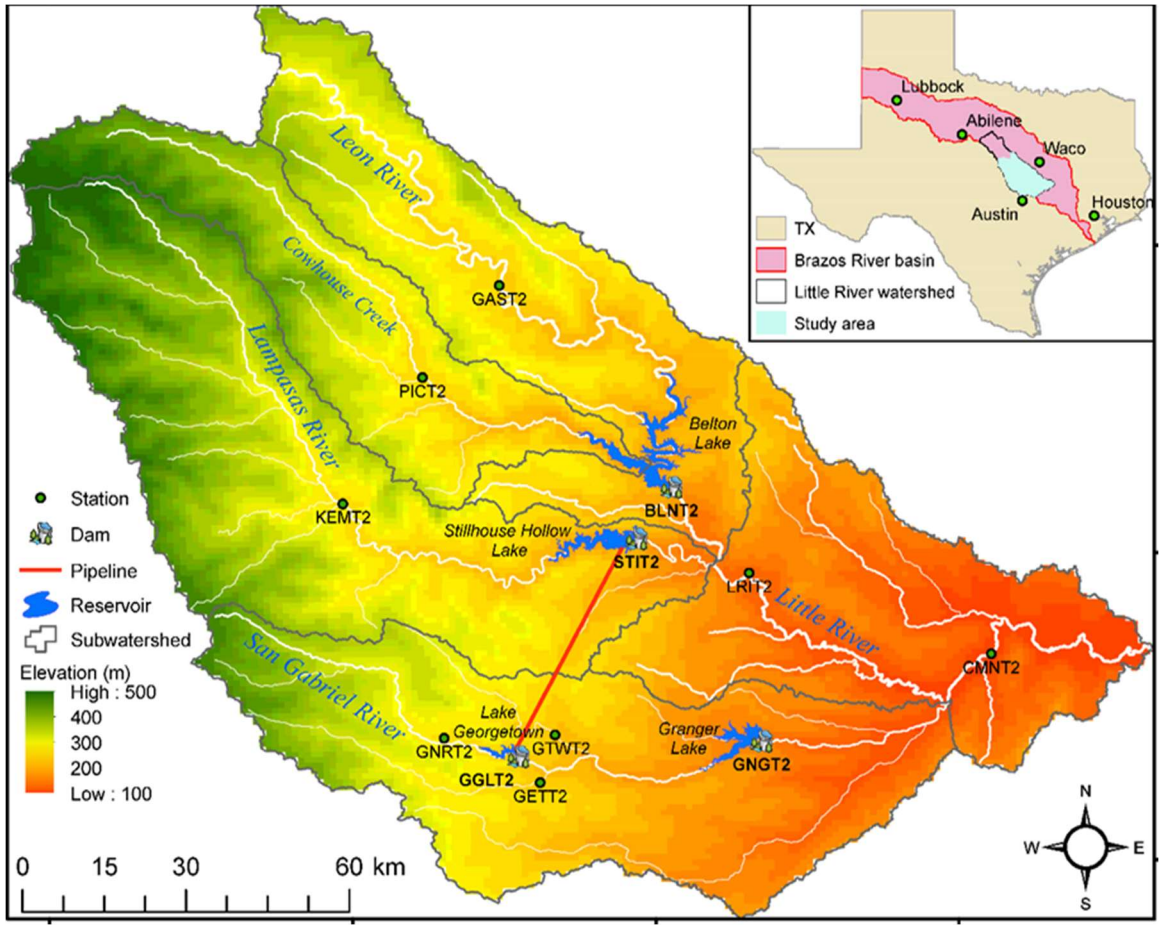

Figure 1-1: The central Texas FIRO Pilot on the Little River System, where four USACE multiuse reservoirs are situated, namely Lake Georgetown, Granger Lake, Stillhouse Hollow Lake, and Belton Lake.

The FIRO Pilot site is home to four USACE multiuse reservoirs: Lake Georgetown, Granger Lake, Stillhouse Hollow Lake and Belton Lake (Table 1-1). Among these, Lake Georgetown has

the smallest storage capacity while serving an area with a rapidly expanding population, making it particularly vulnerable to supply shortfall. To improve water supply reliability, the BRA has constructed a pipeline as part of the Williamson County Regional Raw Water Line. The first phase of the project connects Stillhouse Hollow Lake to Lake Georgetown, and its operation commenced in 2006.

Table 1-1: USACE Multipurpose Reservoirs in Texas FIRO Pilot

Reservoirs	NWS ID	Surface Area (acres)	Conservation Capacity (acre-ft)	Upstream Area (square miles)
Lake Georgetown	GGLT2	1.297	38,005	246
Granger Lake	GRNT2	4,064	51,822	709
Stillhouse Hollow Lake	STIT2	6,429	229,796	1318
Belton Lake	BLNT2	12,385	432,631	3560

Three USGS gauging stations are located upstream of Stillhouse Hollow Lake and Belton Lake where stage height and discharge data are available (Table 1-2).

Table 1-2: USGS gauging stations in the FIRO Pilot with flow data.

Location	USGS ID	NWS ID	Upstream Area (square miles)	Downstream Reservoir
Leon River at Gatesville	08100500	GAST2	2,342	Belton Lake
Cowhouse Creek at Pidcoke	08101000	PICT2	455	Belton Lake
Lampasas River nr Kempner	08103800	KEMT2	818	Stillhouse Hollow Lake

The focus of the beginning phases of the FIRO pilot has been on creating and evaluating the skills of ensemble streamflow forecasts from the NWS Hydrologic Ensemble Forecast Service (HEFS; Demargne, 2014), and on identifying potential opportunities of using these forecasts to guide reservoir operations. HEFS was developed between 2003 and 2010 as a major enhancement to the Ensemble Streamflow Prediction (ESP; Curtis and Schaake, 1979; Day, 1985). ESP has been operational at various River Forecast Centers since the 1970s; it relies solely on historical traces of precipitation (and in certain locations, temperature) as forcing to produce ensemble streamflow forecasts, HEFS has the ability to ingest forecasts from numerical weather Prediction (NWP) models and offers new capabilities such as postprocessing of ensemble streamflow forecasts. As shown in Fig. 1-2, the current architecture of HEFS comprises the following modules:

- Meteorological Ensemble Forcing Processor (MEFP; Wu et al, 2010): postprocessing raw NWP forecasts to produce ensemble traces of forcing variables including precipitation and temperature for each forecast point.
- Hydrologic/reservoir models: producing streamflow forecasts using forcing variables. These include Sacramento Soil Moisture Accounting (SAC-SMA; Burnash 1973), a routing model, reservoir modules, and a snow ablation model (SNOW-17; Anderson 1976)

- Ensemble Postprocessor (EnsPost; Regonda and Seo, 2008): postprocessing ensemble streamflow forecasts by blending with observations.
- Ensemble Verification System (EVS): computing verification statistics on ensemble forecasts

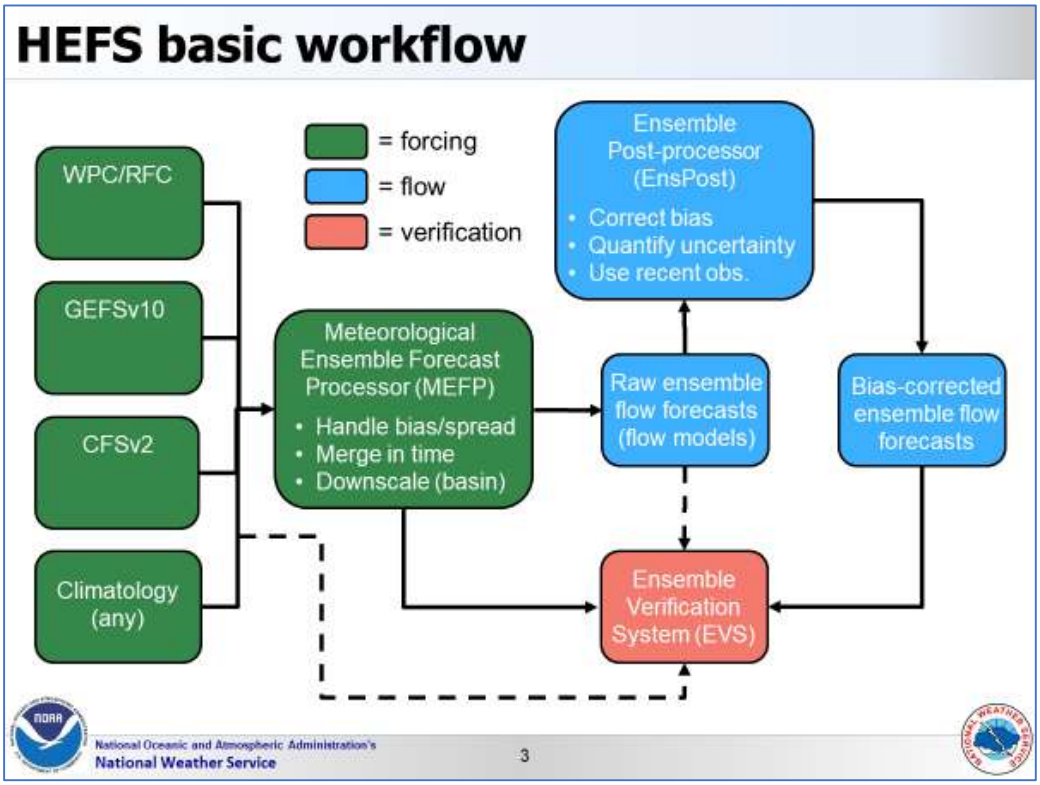

Figure 1-2 Workflow of Hydrologic Ensemble Forecast Service (HEFS). Source: NWS.

The NWS Office of Water Prediction (OWP), which oversees the development of HEFS, has been gradually rolling out the system for different River Forecast Centers (RFCs). At NWS WGRFC, the adoption of HEFS goes back to 2015 with assistance from the research team of Prof. Dong-Jun Seo (Kim et al., 2014), who led the development of HEFS at OWP.

The NWS OWP produced baseline HEFS hindcasts and validation for 2000–2019, a period for which the Global Ensemble Forecast System – version 12 (GEFSv12, please include citation) reforecast is available. However, the HEFS configuration used in producing the hindcasts was based on a slightly older version of the operational forecast system configuration used at the WGRFC that had not integrated more recent changes made at the center, which include updated topology, switching from 6-h to 3-h time step, and integration of updated parameter values from the recent calibration completed in 2022. Therefore, it was necessary to review and update the configuration to be consistent with the current operational settings. In addition, the baseline validation only covers forecast points with USGS observations and excludes those at reservoir inlets. The hindcast configuration of the baseline validation package is summarized in Table 1-3.

Table 1-3: HEFS configuration for OWP baseline validation

Forcing	Initialization Frequency	Time Step	Number of ensemble members	Lead time validated	Locations validated
GEFSv12 precipitation forecasts (day 1–15)	Daily	6-h	41	Day 1–30	GAST2, PICT2, KEMT2
Resampled Climatology (day 16–270)					

2. Project Overview

The present project aims to lay the groundwork for future FIRO investigations at the Texas FIRO Pilot. The specific objectives include:

1. Updating the configurations of Hydrologic Ensemble Forecast Service (HEFS) and reservoir model from NWS WGRFC to be consistent with the latter's real-time river operations and more accurately represent reservoir operations.
2. Expanding the hindcast and validation efforts to inform WGRFC on potential improvements to the configurations to yield more skillful forecasts.
3. Providing forecast data and support to TWDB's Lake Conroe initiative.

The project comprises the following five tasks:

1. Review and revise the hindcast configuration of WGRFC HEFS for the Little River System (LRS).
2. Produce hindcasts for extended lead times using the revised versions of WGRFC HEFS and perform validation.
3. Perform validation on Climate Forecast System-version 2 (CFSv2; Saha et al., 2014) and Global Ensemble Forecast System-version 12 (GEFSv12) subseasonal to seasonal (S2S) precipitation forecasts.
4. Deliver postprocessed ensemble precipitation and streamflow forecasts for HEFS forecast points upstream of Lake Conroe.
5. Report findings to stakeholders through regular meetings.

3. Changes to hindcast configuration for LRS

We compared the model configuration in the baseline validation package against the current WGRFC modeling system and identified two major differences, namely: 1) the operational model now produces forecasts at 3-h rather than 6-h intervals; and 2) the baseline configuration uses values of SAC-SMA parameters prior to the recent calibration concluded in 2022. To maintain consistency with WGRFC operational model, we created three HEFS stand-alone configurations that produce forecasts at 3-h intervals, and incorporated SAC-SMA parameter

values from the latest round of calibration for the LRS completed in 2022 (Lynker Tech, 2022). The main differences among the configurations are the forcing data. The first configuration, labelled as "GEFS-Climatology", mimics the baseline validation that uses GEFSv12 medium-range QPF for the lead time of Day 1–14, and resampled climatology beyond that range. The second configuration, GEFS-S2S uses GEFSv12 subseasonal forecasts that are issued weekly (rather than daily as in GEFS medium-range forecasts). The third configuration, GEFS-CFSv2, uses CFSv2 seasonal forecasts in lieu of resampled climatology.

Table 3-1: Hindcast configurations created for the current project

Configuration Identifier	Day 1-14	Day 15-270	Frequency of Initialization	Lead Time Range
Climatology	Resampled climatology	Resampled climatology	Daily	Day 1-270
GEFS-Climatology	GEFSv12 medium-range	Resampled climatology	Daily/Weekly	Day 1-90/ Day 1-270
GEFS-S2S	GEFSv12 subseasonal	GEFSv12 subseasonal to day 35 and resampled climatology onward	Weekly	Day 1-60
GEFS-CFSv2	GEFSv12 subseasonal	CFSv2	Monthly	Day 1-270

Among these configurations, GEFS-S2S is created to determine the potential skills in GEFS subseasonal forecasts that can be leveraged to benefit forecasts of streamflow anomalies at the S2S range, i.e., week 3–5). GEFS-CFSv2 is created to identify skills in CFSv2 forecasts relative to baseline, resampled climatology for impactful events, e.g., droughts and flooding, at the extended range (beyond 14 days).

4. Evaluation of HEFS hindcasts from new configurations

Hindcasts produced using each configuration listed in Table 3-1 underwent evaluation. The evaluation covers not only precipitation and streamflow, but also reservoir inflow to and pool level at each of the four reservoirs. In addition, the evaluation covers both medium and extended ranges as shown in Table 3-1. Metrics of the evaluation include Continuous Ranked Probability Score (CRPSS) and its Skill Score (CRPSS), Brier Score (BS) and Brier Skill Score (BSS), Reliability diagram, Receiver Operating Characteristic (ROC), and bias. Definitions of these metrics are provided below.

$$CRPS\ (F, x) = \int_{-\infty}^{\infty} (F(y) - I(y - x))^2\, dy$$

Where F(y) is the cumulative density function (CDF) of predictive distribution; y is the threshold value for the forecast variable x; and I(y-x) is the Heaviside function that takes 1 once y exceeds x and 0 otherwise.

CRPSS is the complement of the ratio of the CRPS for the ensemble forecast being evaluated, $CRPS_{forecast}$, and the climatological distribution $CRPS_{climat}$.

$$CRPSS = 1 - \frac{CRPS_{forecast}}{CRPS_{climat}}$$

Brier Score (BS) is defined as the mean difference between forecast probability and occurrence of an event, namely,

$$BS\ (F, x) = \frac{1}{N}\sum_{i=1}^{N}(F(y) - I(y - x))^2$$

Brier Skill Score is defined to gauge the skills of an ensemble forecast suite relative to a reference forecast suite, e.g., climatology, i.e.:

$$BSS = 1 - \frac{BS_{control}}{BS_{reference}}$$

Where $BS_{control}$ is and $BS_{reference}$ are Brier Score for the control and reference forecast suites. A reliability diagram is a plot of frequency of observations above a given threshold within a subsample, for which the forecast probability exceeding (or below) the threshold equals a prescribed value. It characterizes under or overconfidence in the forecast.

Receiver Operating Characteristic (ROC) is a plot of true positive rate (TPR) against false alarm rate (FAR) for a prescribed threshold. It characterizes discrimination skills of forecasts. Originally developed for deterministic forecasts, it can be adapted to ensemble forecasts by averaging the TPR and FAR for each ensemble member. Just like BSS, a ROC score can be defined for a control forecast suite against a reference forecast:

$$ROC\ score = 1 - \frac{ROC_{control}}{ROC_{reference}}$$

The streamflow forecasts at forecast points collocated with USGS stations are verified using USGS flow records, whereas those at reservoir inlets are verified using reservoir inflow series constructed by the USACE-SWF with a water balance approach. In this approach, the water level is assumed uniform across the entire reservoir at each instant.

$$Q_I(t) = A(t)\frac{\Delta h}{\Delta t} + Q_o(t)$$

Where Q_I is the incremental inflow volume (including direct rainfall on reservoir); Q_o is the loss term that comprises reservoir release, evaporative loss, and withdrawal; A is the reservoir area; h is the reservoir water level and t is time.

In the following subsections, we will first present the outcomes of evaluations for each HEFS configuration.

4.1. GEFS-Climatology:

The HEFS hindcasts produced using GEFS-Climatology configuration were evaluated for the entire 20-year period from 2000 to 2019. We first summarize the performance of HEFS forecasts using metrics including CRPSS, BSS, reliability diagrams, ROC scores, and percentage biases. Then we compare the performance of streamflow **simulations** by NWS hydrologic model using the previous and updated SAC-SMA parameter values. Finally, we present a case study to illustrate the skills of ensemble forecasts for an impactful flood event and highlight areas for improvements.

Overall Forecast Skills

The CRPSS is first computed for ensemble streamflow forecasts produced using the GEFS-Climatology at the seven forecast points described in Section 3. The results for the three USGS sites, namely GAST2, PICT2, and KEMT2 are shown in Figures 4-1 – 4-4 for 1–9-month lead.

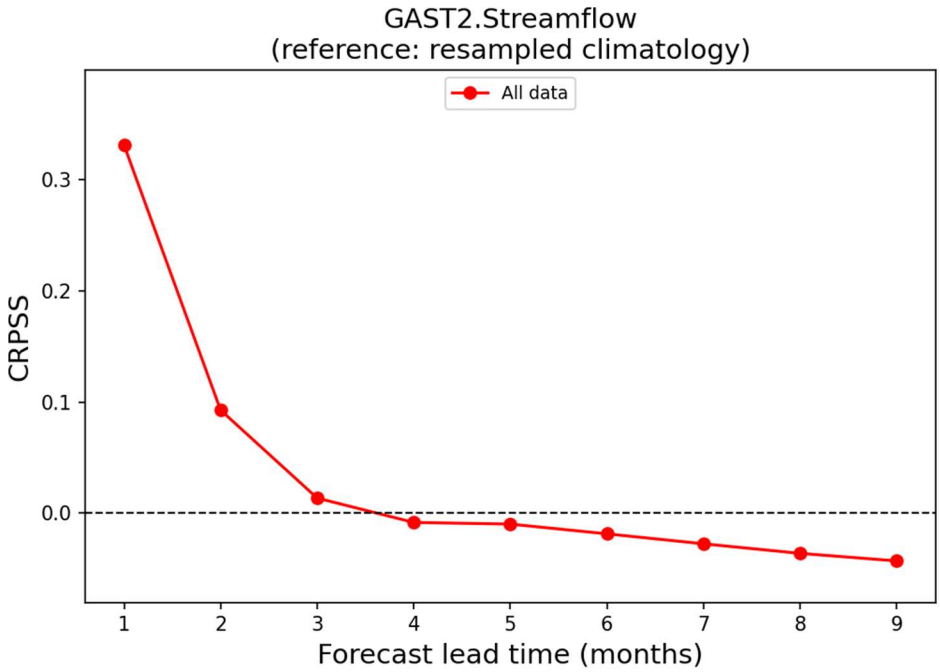

Figure 4-1: CRPSS for HEFS ensemble streamflow forecasts produced using the GEFS-Climatology at Leon River at Gatesville (GAST2).

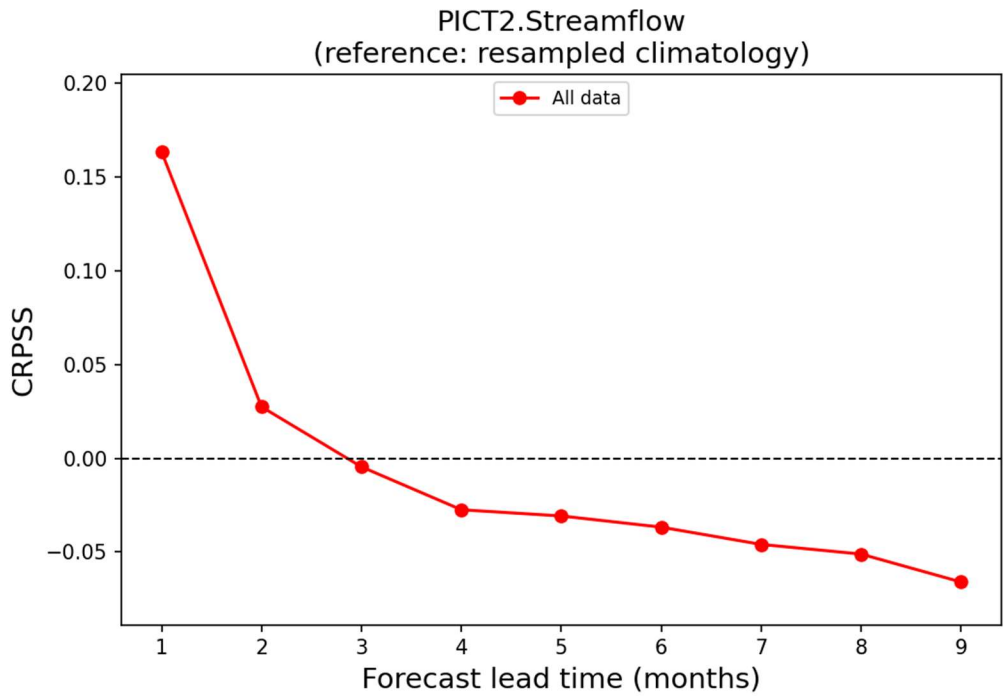

Figure 4-2: As Figure 4-1, except at PICT2.

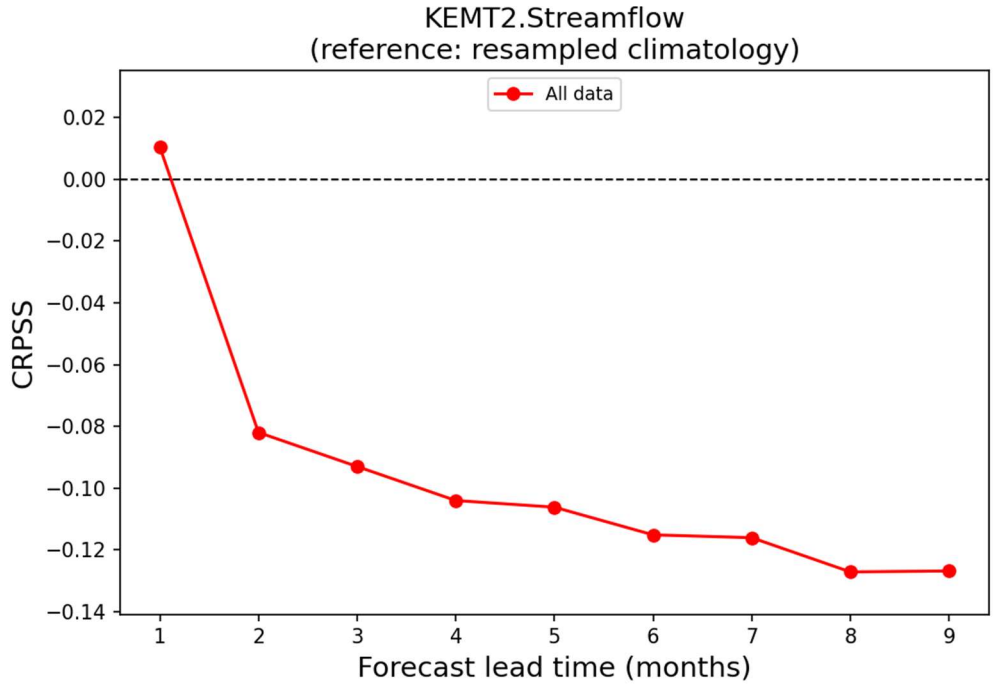

Figure 4-3: As Fig. 4-1, except at KEMT2.

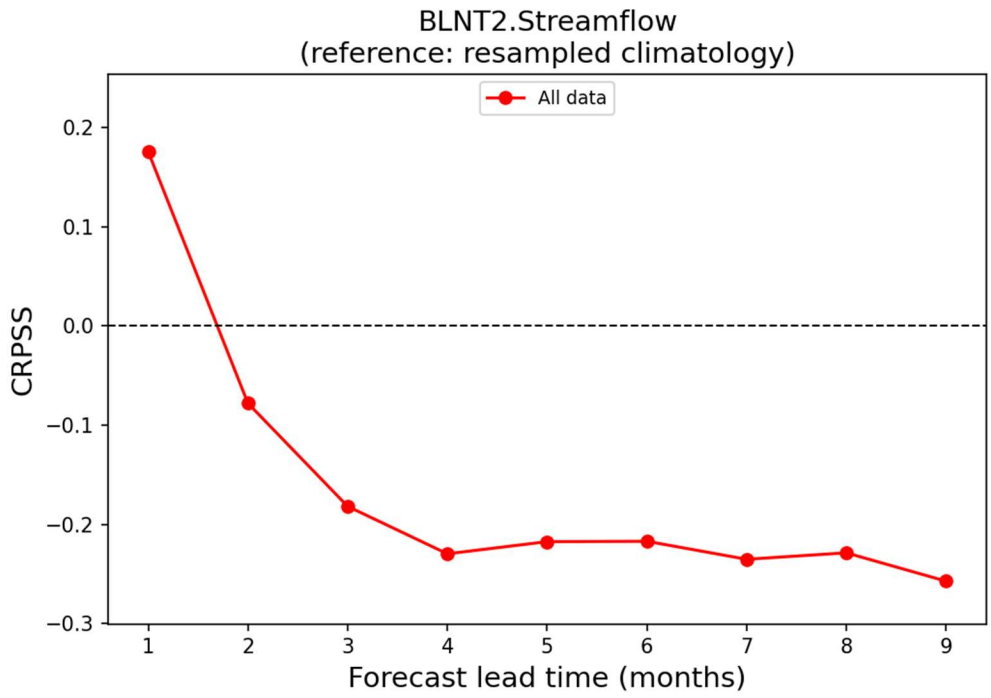

Figure 4-4: CRPSS for HEFS ensemble inflow forecasts produced using the GEFS-Climatology to Belton Lake (BLNT2), and verified against USACE daily reconstructed inflow.

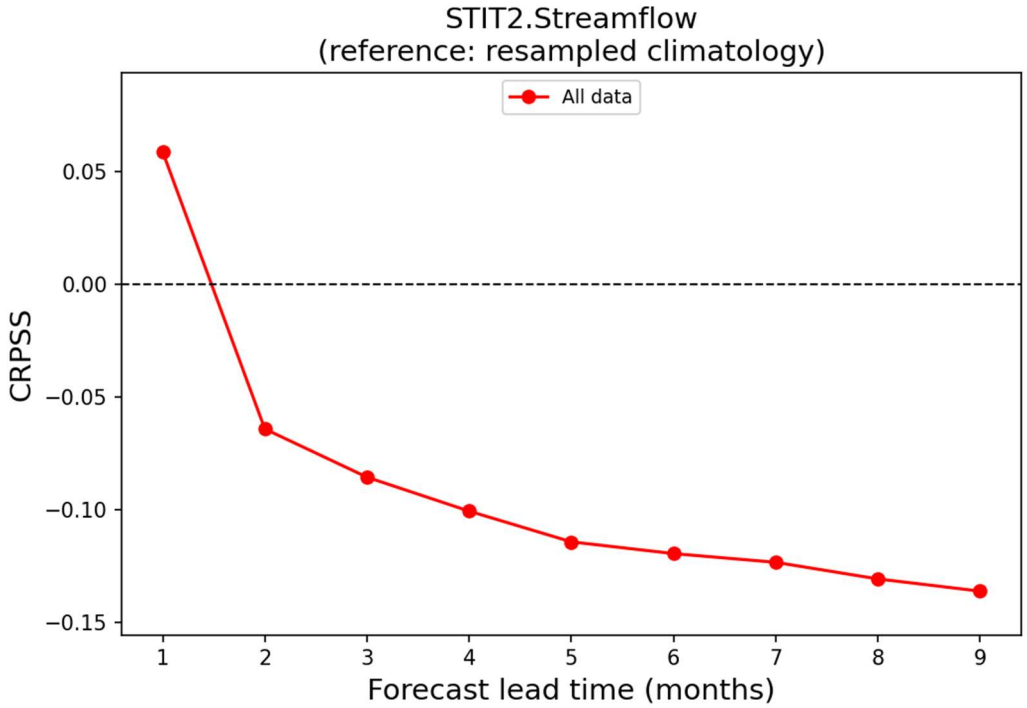

Figure 4-5: As Fig. 4-4, except for inflow to Stillhouse Hollow (STIT2) and verified against USACE daily reconstructed inflow.

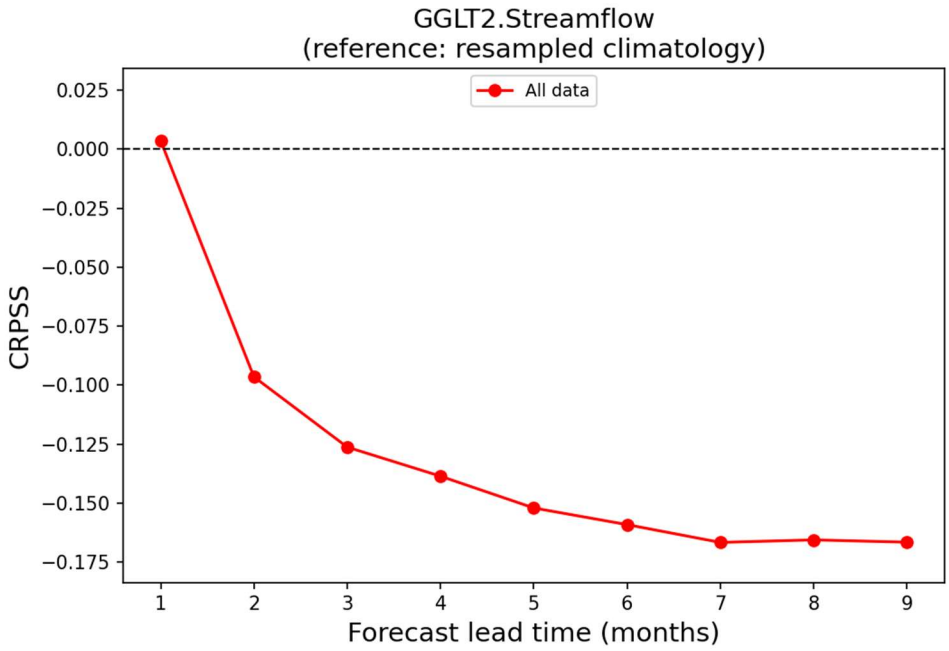

Figure 4-6: As Fig. 4-4, except for inflow to Lake Georgetown (GGLT2), and verified against USACE daily reconstructed inflow.

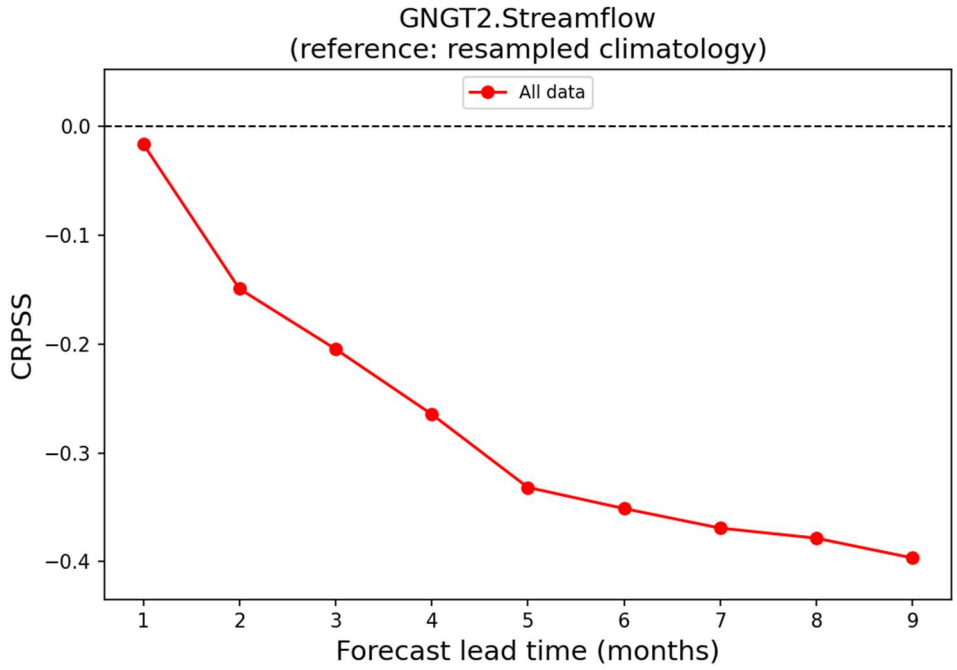

Figure 4-7: As Fig. 4-4, except for inflow to Lake Granger (GNGT2), and verified against USACE daily reconstructed inflow.

These results show that the ensemble streamflow forecasts are generally more skillful than climatology for the lead time range of 1–3 months. Among the three forecast points, HEFS forecasts are the most skillful at GAST2 (to month 3), followed by PICT2 (month 2) and KEMT2 (month 1). Note that GAST2 features the largest upstream drainage area (2342 mi^2), which implies longer travel time that extends the predictive skills in precipitation forecasts. By comparison, PICT2 and KEMT2 are associated with smaller drainages, and KEMT is known for its flashness in response that limited the time horizon of skillful streamflow forecasts.

The results at the four reservoir inlets largely mirror those at the upstream forecast points, except that it is noted that the CRPSS values for inflows at GGLT2 (Georgetown) and GNGT2 (Granger) are systematically lower than those at BLNT2 (Belton) and STIT2 (Stillhouse Hollow). At GNGT2, CPRSS value even at 1-month lead is beneath zero, indicating a lack of skills relative to resampled climatology. The contrasts can be attributed to the following factors: 1) both Georgetown and Granger feature smaller upstream drainage areas, and a flashier response to rainfall, than Belton and Stillhouse Hollow, and 2) release from Georgetown is a major source of inflow to Granger, and forecast skills of the release are hampered by inaccuracies in water balance components, including withdrawal, evaporation, and inflow, as prescribed to or represented by the reservoir module used in HEFS.

The BSS values computed at 10, 90 and 99% quantiles of monthly flows at the seven forecast points are shown in Figs. 4-8 – 4. 14.

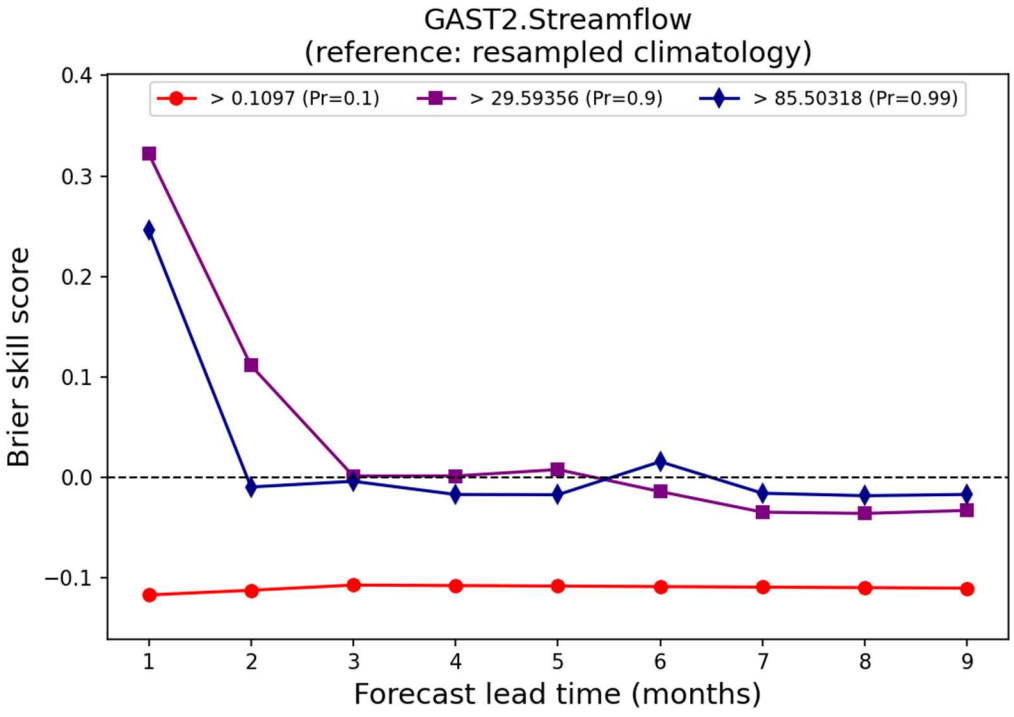

Figure 4-8: BSS of HEFS ensemble streamflow forecasts versus lead time at GAST2. The skill score is computed using streamflow forecasts and observations averaged onto monthly intervals at 10, 90 and 99% quantile thresholds derived from observations.

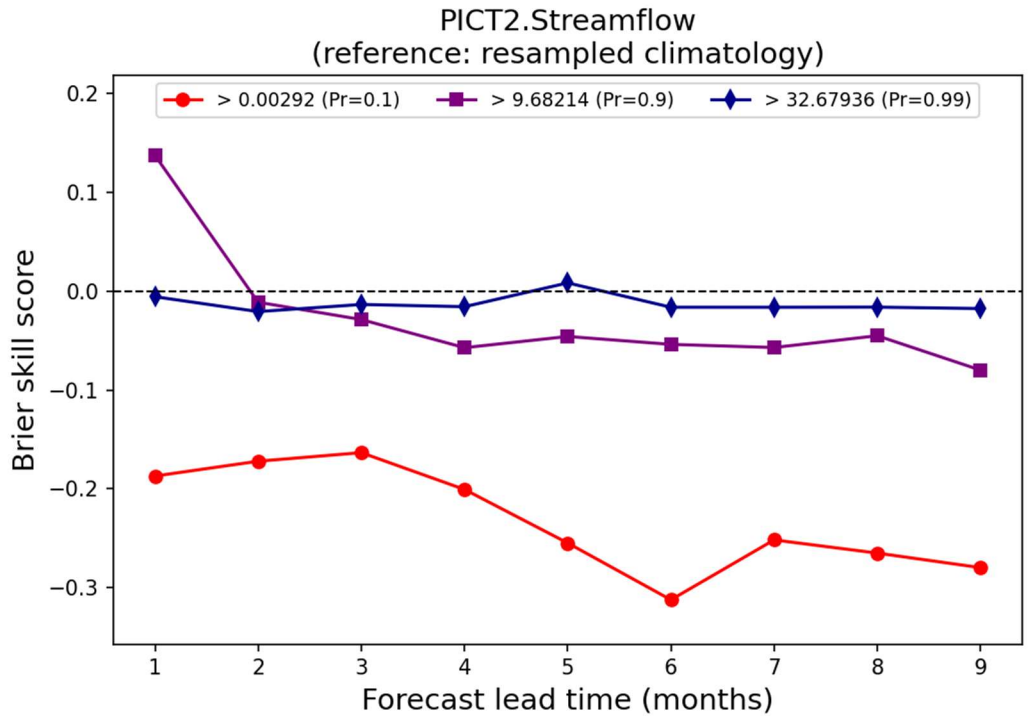

Figure 4-9: As Fig 4-8, except at PICT2.

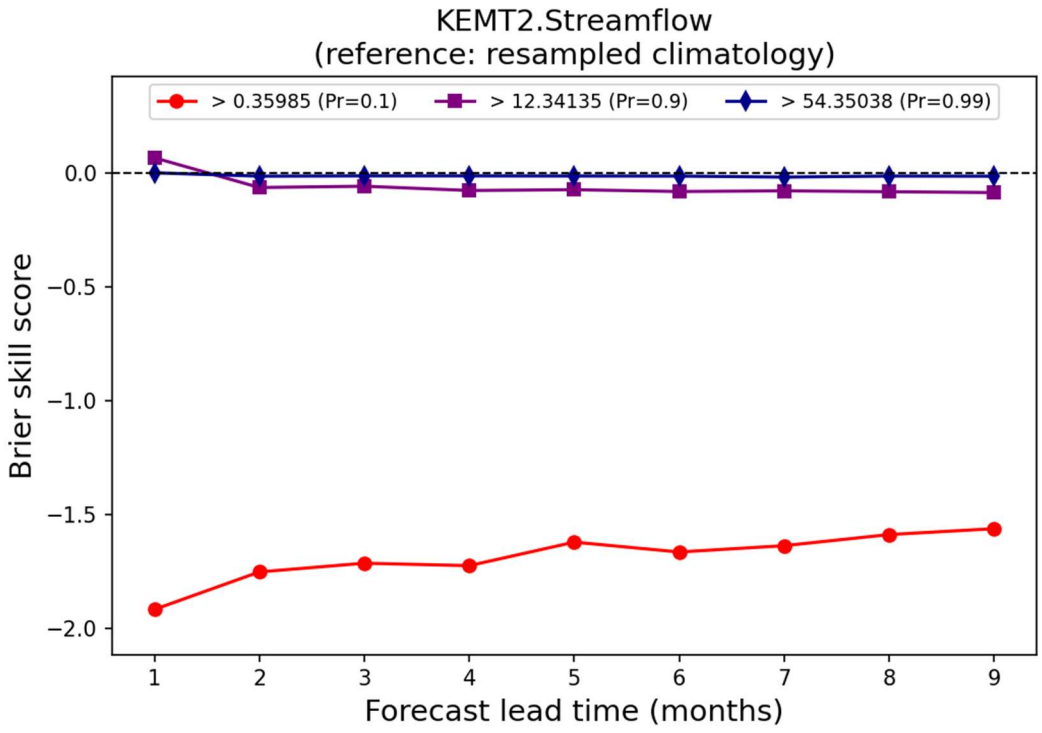

Figure 4-10: As Fig 4-8, except at KEMT2.

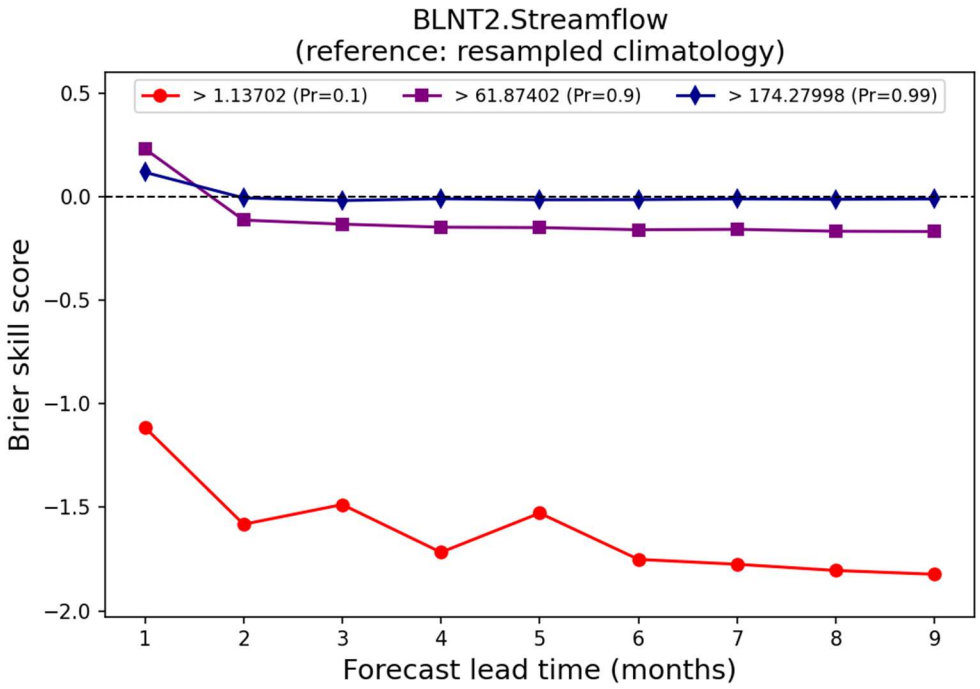

Figure 4-11: As Fig 4-8, except for inflow to Lake Belton (BLNT) and verified against USACE daily reconstructed inflow.

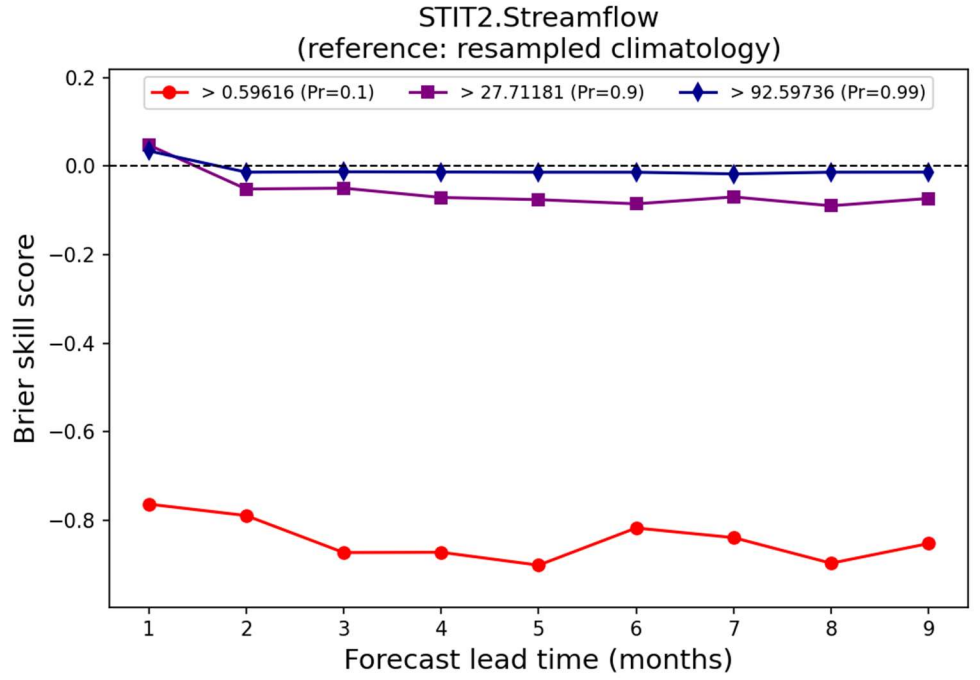

Figure 4-12: As Fig 4-8, except for inflow to Stillhouse Hollow Lake (STIT2) and verified against USACE daily reconstructed inflow.

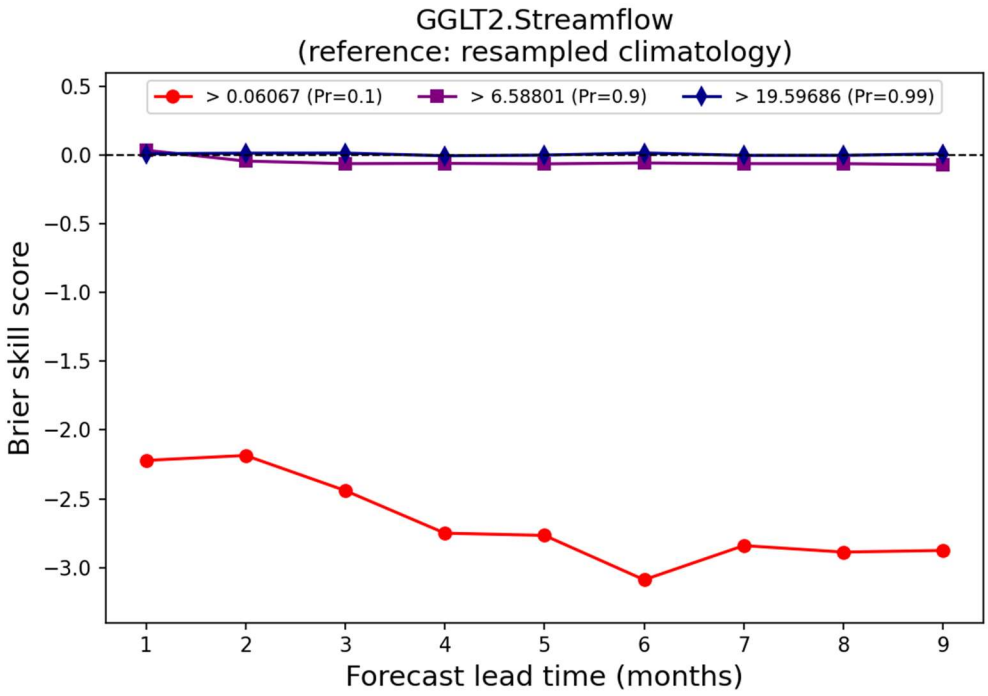

Figure 4-13: As Fig 4-8, except for inflow to Lake Georgetown (GGLT2) and verified against USACE daily reconstructed inflow.

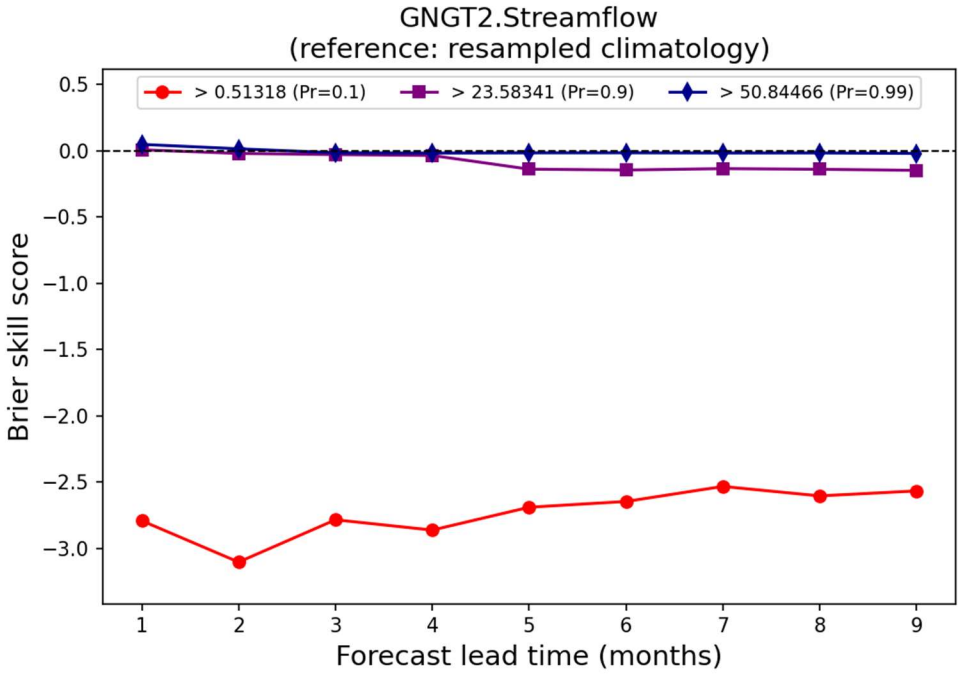

Figure 4-14: As Fig. 4-8, except for inflow to Lake Granger (GNGT2) and verified against USACE daily reconstructed inflow.

Key observations from the BSS plots at the three upstream forecast points include:

- Skills are positive for moderate/high flow thresholds at 1-month lead at GAST2, the forecast point with the largest drainage area.
- Skills are positive for moderate flow threshold at 1-month lead at PICT2 and KEMT2, though only marginally so at the latter. There is no skill at the highest threshold (99% quantile) at the two sites.
- There is no skill at the low flow threshold (10% quantile) at all three sites.

The observations at the four reservoir inlets again resemble those at the upstream forecast points. Among the four sites, BSS values at the moderate (90%) and top (99%) thresholds are barely above zero at Georgetown and Granger even at the 1-month lead, echoing the plots of CRPSS where the skills appear much lower at the two sites. At all four sites, the HEFS forecasts exhibit lower skills than climatology in forecasting low flows (10% quantile).

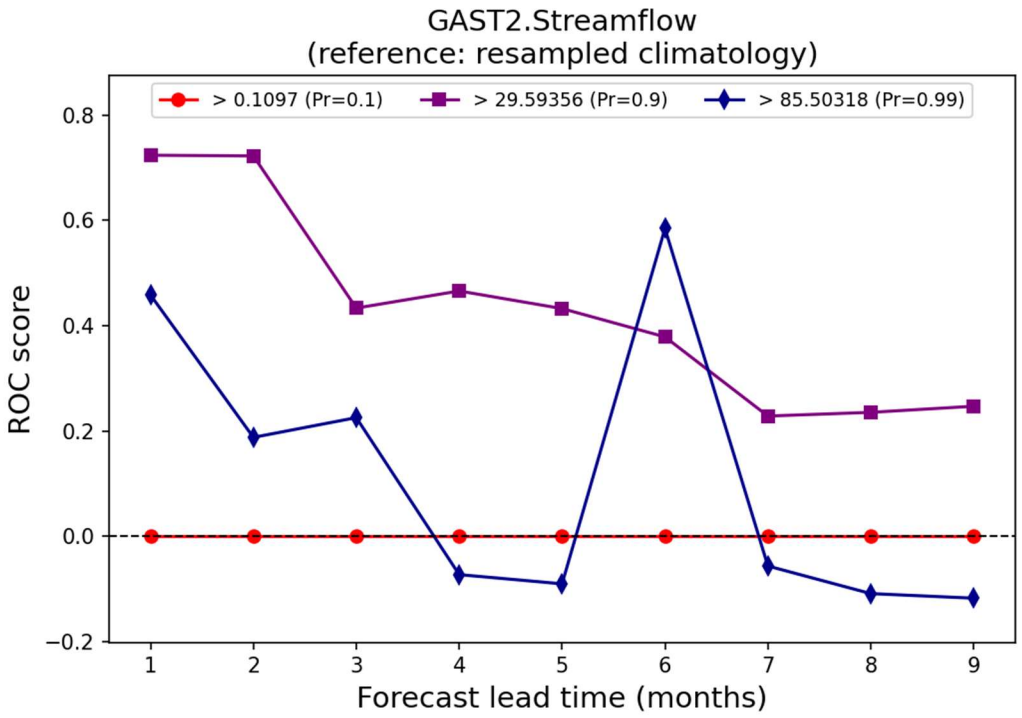

Figure 4-15: ROC of HEFS forecasts with GEFS-Climatology computed at three thresholds, i.e., 10, 90 and 99% quantiles, at GAST2.

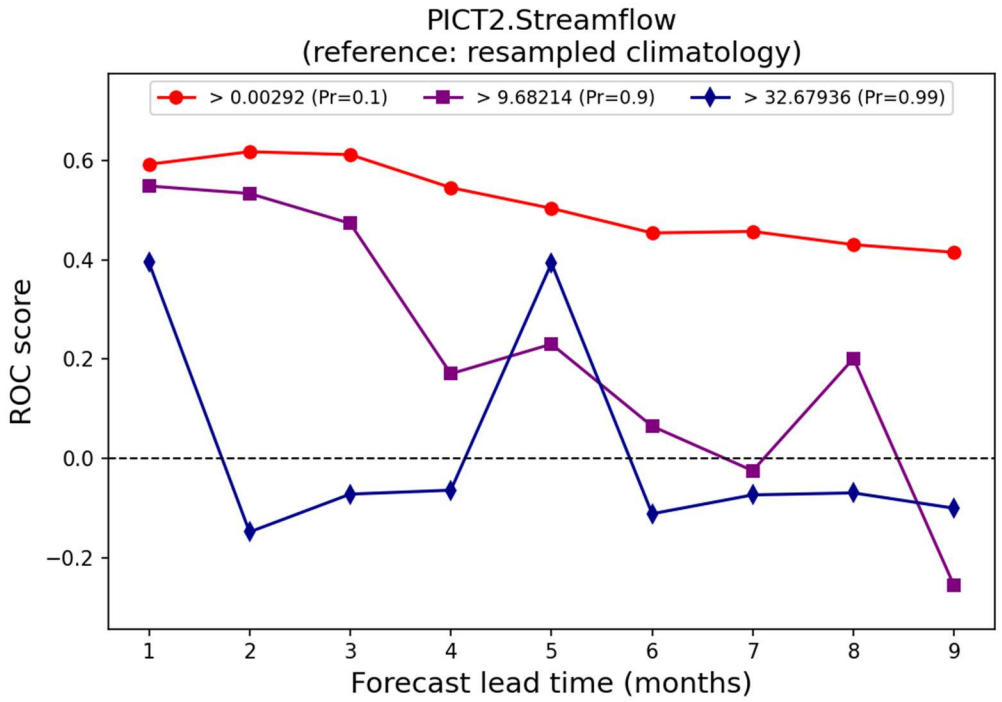

Figure 4-16: As Fig. 4-15, except at PICT2.

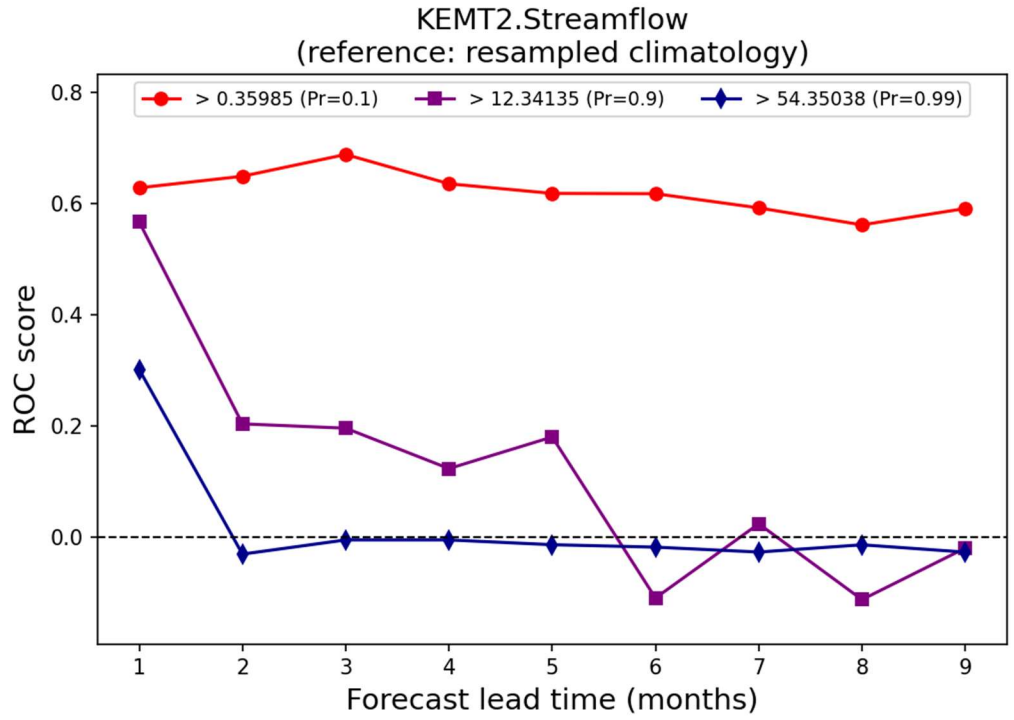

Figure 4-17: As Fig. 4-15, except at KEMT2.

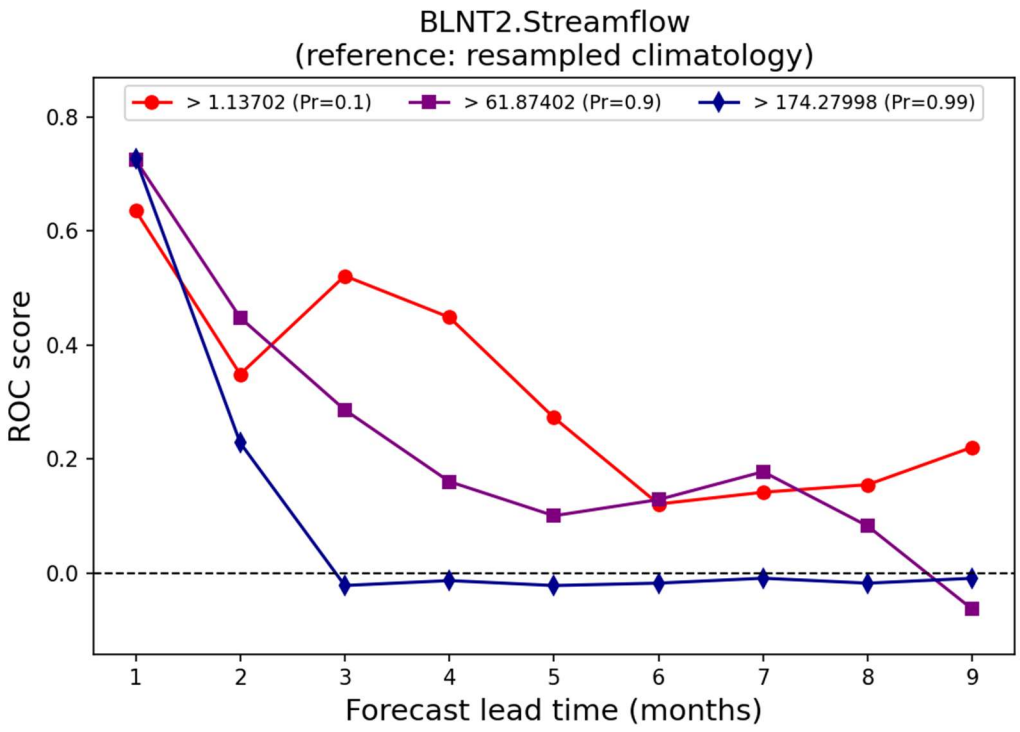

Figure 4-18: As Fig. 4-15, except for inflow to Lake Belton and verified against USACE reconstructed inflow.

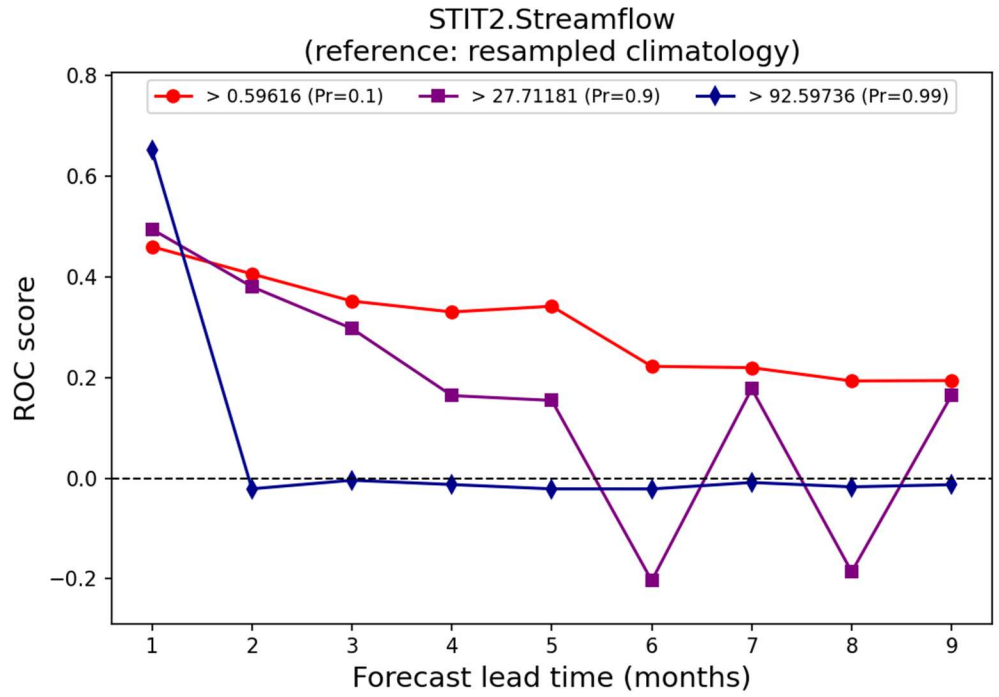

Figure 4-19: As Fig. 4-15, except for inflow to Stillhouse Hollow Lake (STIT2).

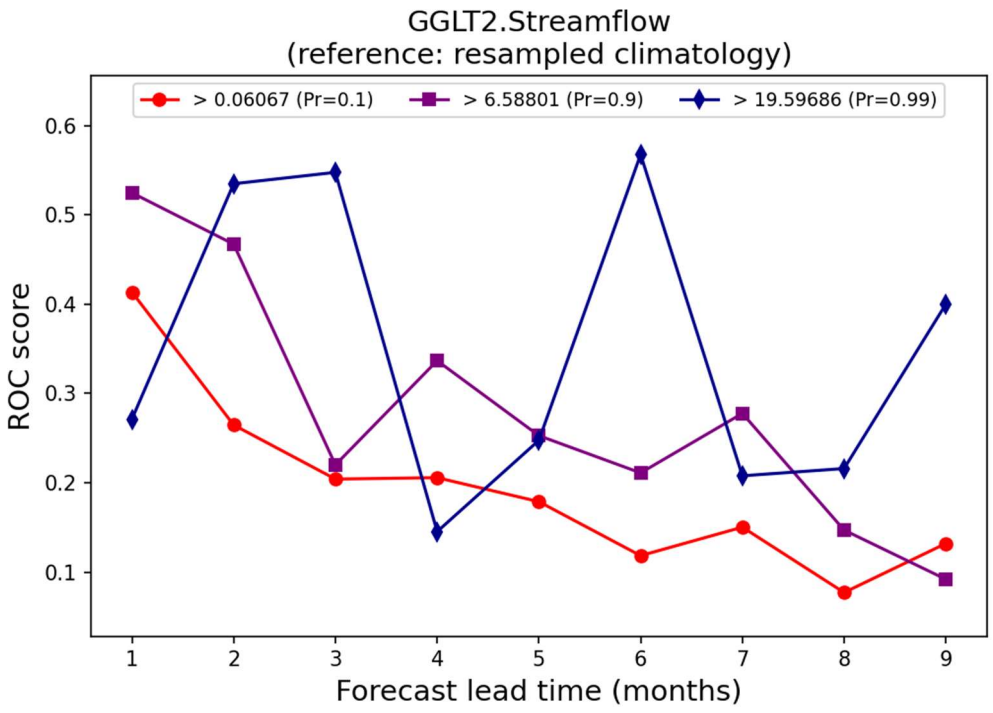

Figure 4-20: As Fig. 4-15, except for inflow to Lake Georgetown (GGLT2).

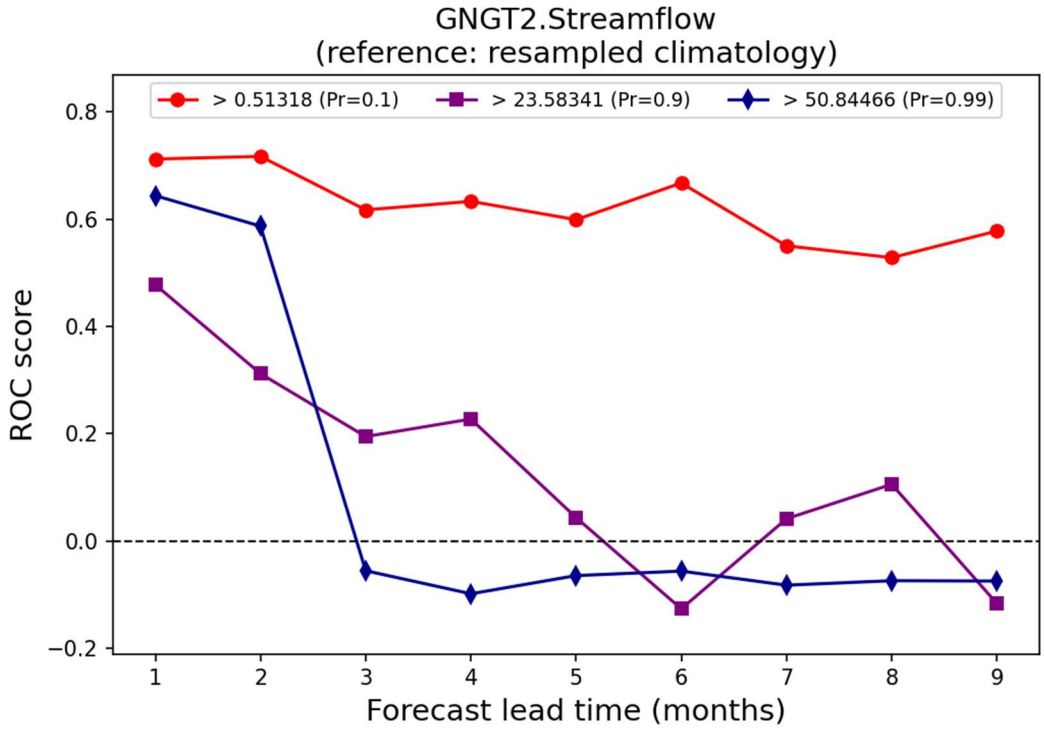

Figure 4-21: As Fig. 4-15, except for inflow to Granger Lake (GNGT2).

ROC scores provide a complementary perspective on the discrimination skills of forecasts, and the ROC scores for monthly flows computed at each forecast points are shown in Figs. 4-15 – 4-21.

Notable observations at the three upstream points are summarized below:

- There are discernible discrimination skills for moderate high flow thresholds at lead times well beyond the first month. At GAST2, the lead time with positive ROC scores extend to 9 months, whereas at PICT2 and KEMT2, it extends to 6 and 5 months, respectively.
- Discrimination skills for the high flow threshold (99% quantile) are generally lower. For GAST2, these skills extend to 3-month lead time, whereas for the smaller watersheds the skills are confined to shorter lead times (1 month).

The ROC scores of ensemble inflow forecasts vary widely among the four reservoirs. For Lake Belton (BLNT2), ROC scores for moderate and high flow thresholds are consistently above zero to 8 or 9-month lead. At the low flow threshold, the scores are positive for the first two months. By contrast, for Stillhouse Hollow Lake, ROC scores for the high flow thresholds are positive till 2-month lead, whereas for the moderate and low flow thresholds the scores stay positive at much longer time horizons (3 months and 9 months, respectively). For Lake Georgetown, the scores for all thresholds stay positive through the entire forecast horizon (1–9 months), whereas at Granger, the scores exhibit similar declining patterns as seen at Stillhouse Hollow. The variable magnitude and lead-time dependence of ROC scores contrasts sharply with that of BSS, suggesting mixed skills of ensemble streamflow forecasts in capturing lower to moderate/high flows.

It should be noted that, unlike the streamflow at the upstream forecast points, the verification of reservoir inflow was subject to several constraints. First, at the time the evaluation was performed, the NWS hydrologic model, namely the Sacramento Soil Moisture Accounting (SAC-SMA) had not been calibrated for the ungauged portions of the drainage upstream of each reservoir, and this may limit the skills of inflow forecasts. Second, there are considerable uncertainties in the inflow estimates that serve as the verification reference. As indicated earlier, these estimates were constructed using the water balance method that relies on several assumptions, including the uniform water level (level-pool) assumption that may not be valid during flood events or under windy conditions.

The skills of HEFS postprocessed precipitation and streamflow forecasts are further illustrated through a set of statistics computed at daily increments. These are shown in Figs. 4-22 – 4-24.

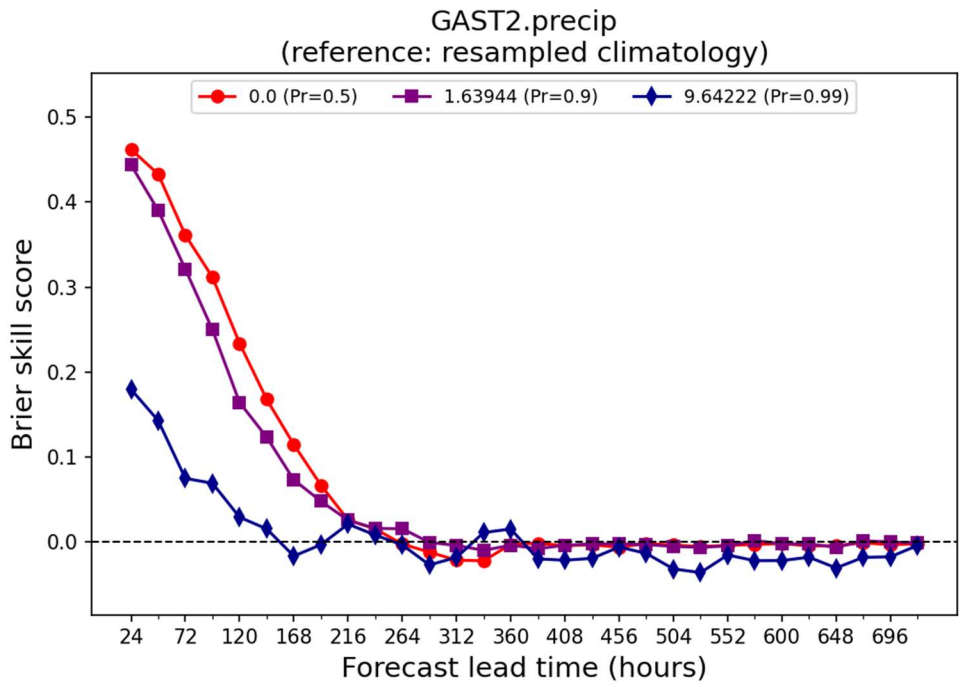

Figure 4-22: BSS of HEFS ensemble precipitation forecasts versus lead time at GAST2 at the 50, 90 and 99% quantile thresholds. The skill score is computed using forecasts and observations averaged onto daily intervals against resampled climatology.

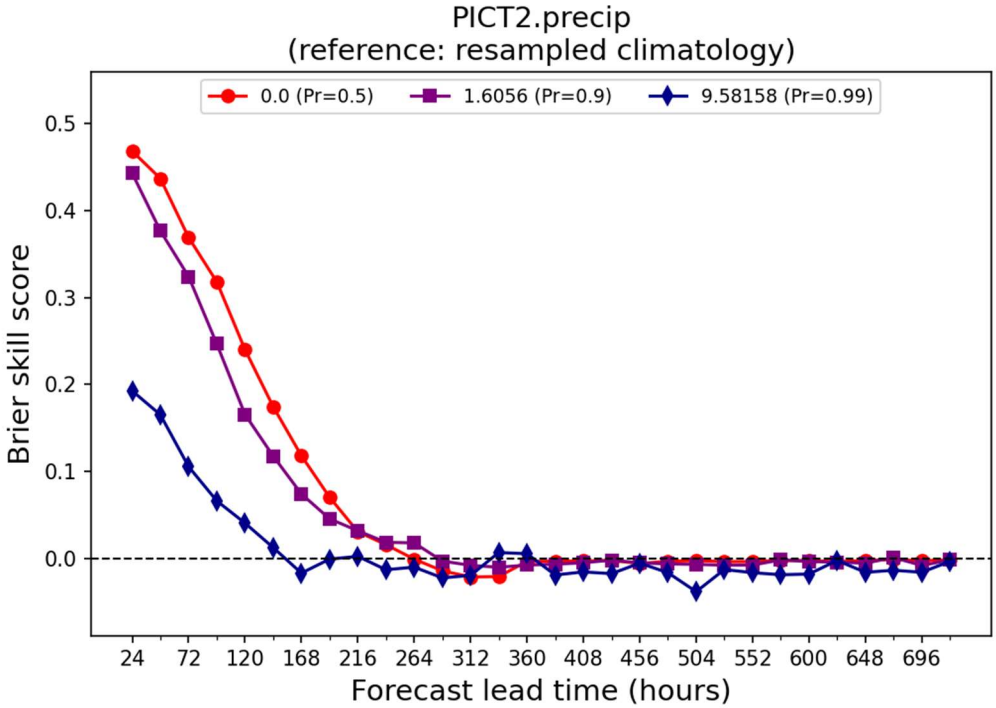

Figure 4-23: As Fig. 4-22, except at PICT2.

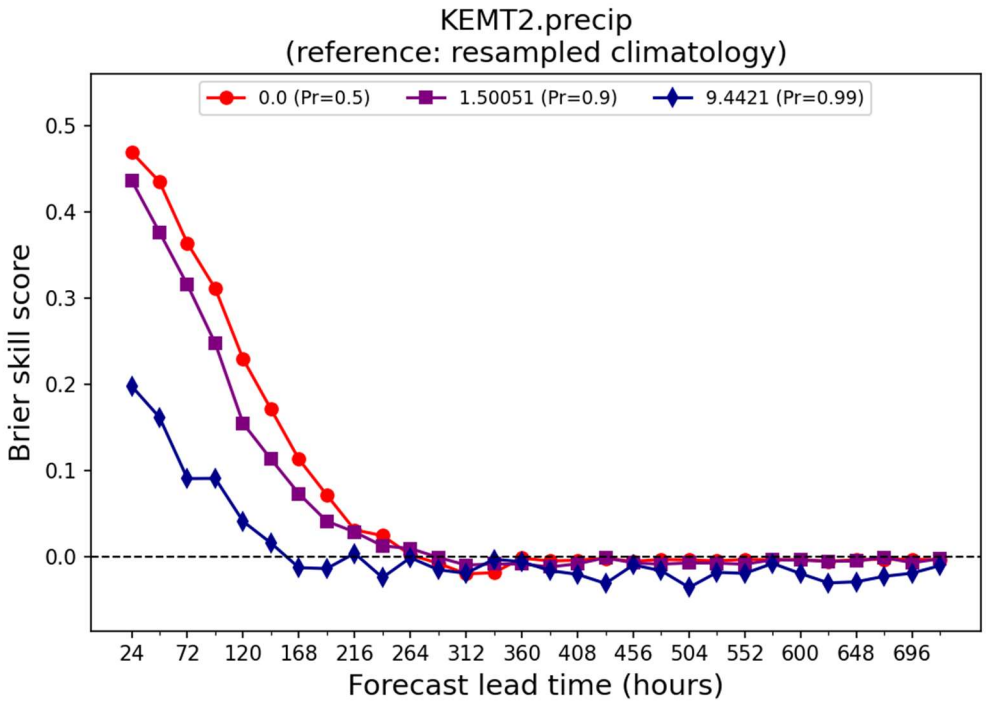

Figure 4-24: As Fig. 4-22, except at KEMT2.

The BSS values of daily ensemble precipitation forecasts are largely similar at the three locations. The following observations are evident:

- Forecast skills decline towards higher precipitation thresholds.
- Ensemble precipitation forecast for the lower (50%) and moderate (90%) quantile thresholds are more skillful than climatology till about day 12 (hour 288).
- Ensemble precipitation forecast for the highest (90%) quantile thresholds are more skillful than climatology till about day 6 (hour 144).

As shown in Figs. 4-25 to 4-27, the ROC scores of daily ensemble precipitation forecasts suggest that the forecasts are skillful to around day 13 irrespective of thresholds and at all locations. In addition, the scores are practically indistinguishable at the moderate and higher thresholds.

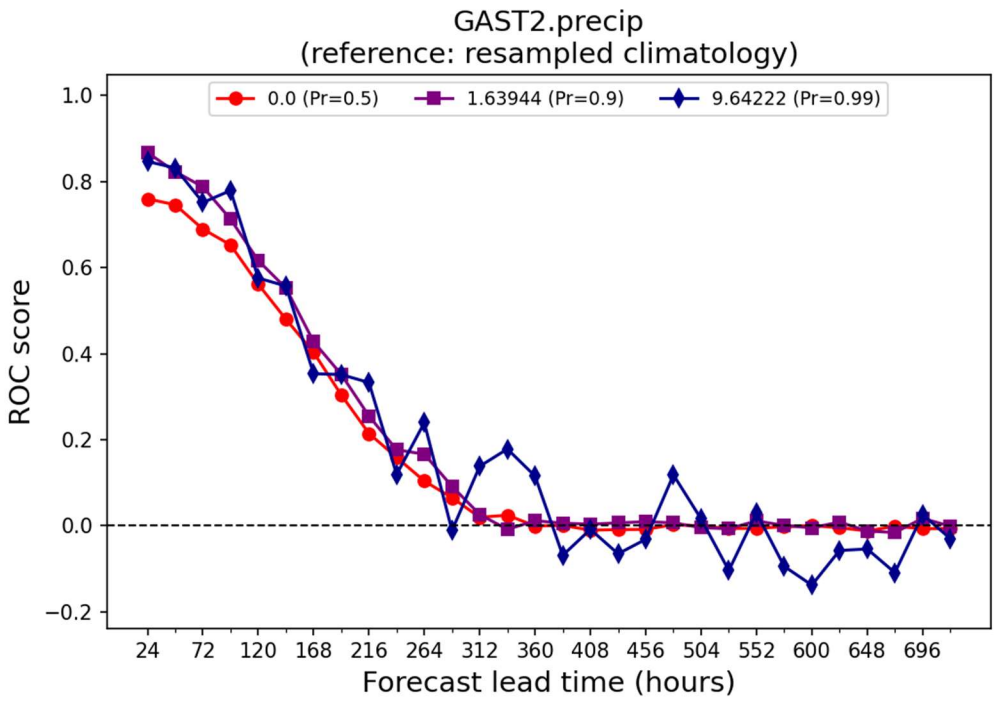

Figure 4-25: ROC score of HEFS ensemble precipitation forecasts versus lead time at GAST2 at the 50, 90 and 99% quantile thresholds. The skill score is computed using forecasts and observations averaged onto daily intervals against resampled climatology.

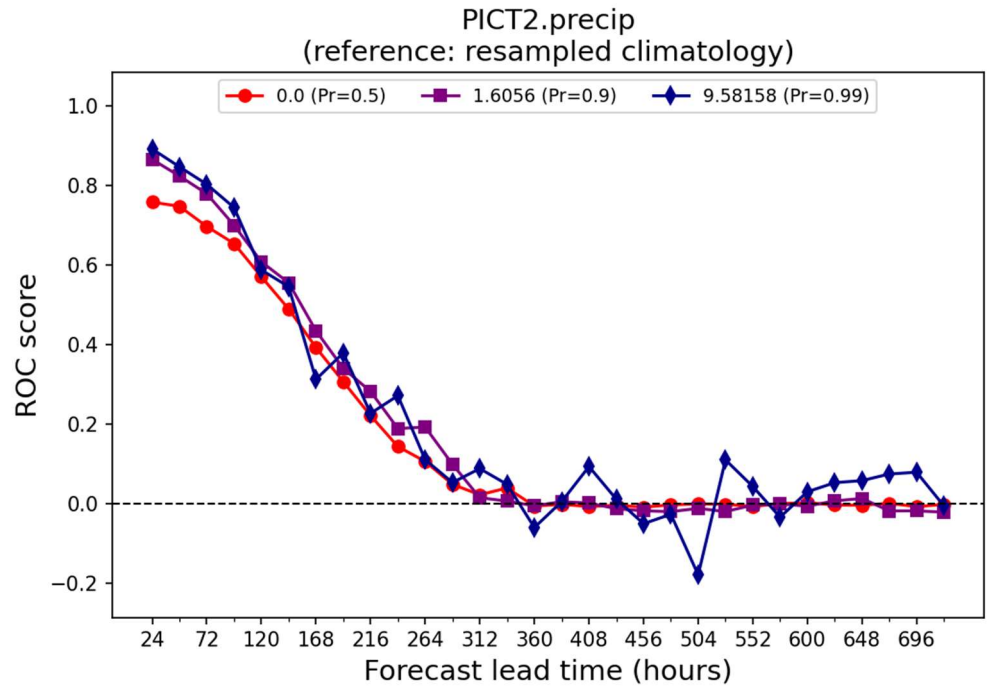

Figure 4-26: As Fig. 4-25, except at PICT2.

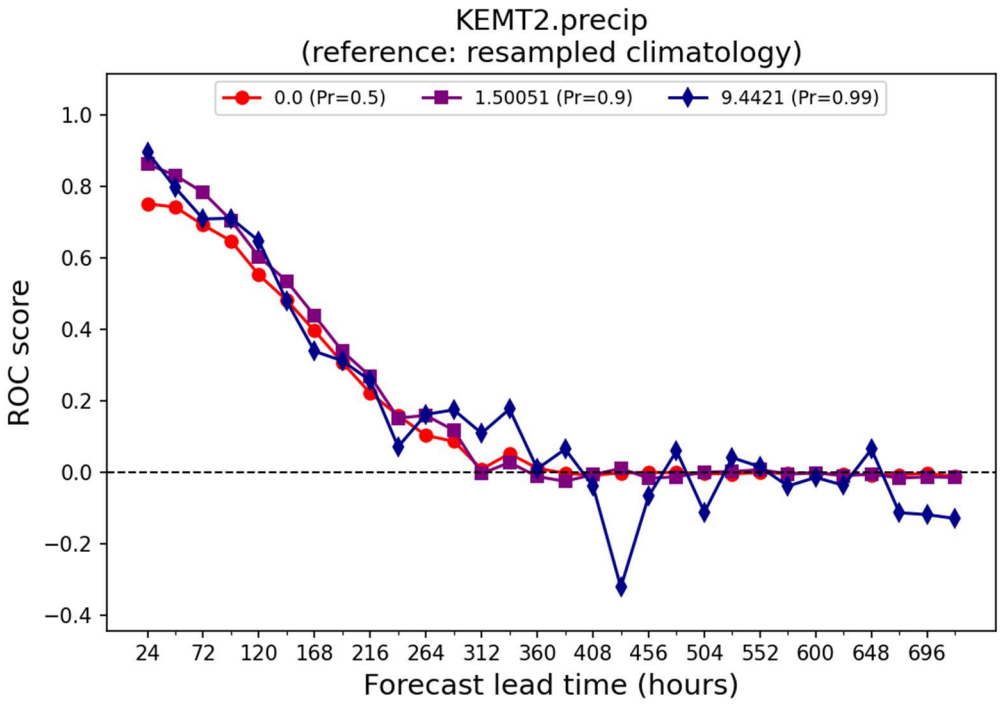

Figure 4-27: As Fig. 4-25, except at KEMT2–.

The verification statistics of daily ensemble streamflow forecasts at the three USGS stations, including BSS and ROC scores, are shown in Figs 4-28 – 4-33. Note that these scores are computed using similar streamflow forecasts forced by resampled climatology. Notable observations include:

- Except for KEMT2, BSS of ensemble streamflow forecasts tends to be the highest at the moderate threshold (90%), and lower at the top and the bottom thresholds (i.e., 90%9 and 50%).
- Across lead times, BSS tends to be higher on days 3–8, rather than on day 1.
- At the top threshold (90%), BSS stays above zero until days 12–13, pointing to skillfulness of GEFS forecasts relative to climatology within this range. Among the three sites, BSS declines at a slower pace at GAST2, which features the largest drainage area.
- At the bottom threshold (50%), BSS remains positive till day 30 without a conspicuous declining trend with lead time, though its values tend to be low.
- The ROC scores tend to be the highest at the top threshold (90%), and the lowest at the bottom threshold (50%).
- The lead time-dependent ROC scores differ widely across the three thresholds and among the three sites.
    - At all three thresholds, ROC scores exhibit an upward trend at shorter lead times. At the moderate and bottom thresholds (50% and 0.90, respectively), the ROC scores increase with lead times from day 1 to day 7 or 10. For the middle threshold (90%0), ROC scores decline onward at longer lead times, whereas the

trend for the bottom threshold is not clear. At the top threshold (90%), the ROC scores increase until days 4–5 and decline onward.

- The skills in forecasting exceedance of bottom/middle thresholds, at least for the shorter lead times (< day 10), tend to be higher at GAST2, which is associated with the largest drainage, and the lower at PICT2 and KEMT2 which feature smaller catchment areas. At KEMT2, the ROC scores for the middle threshold are negative till day 5 (Fig. 4-33).

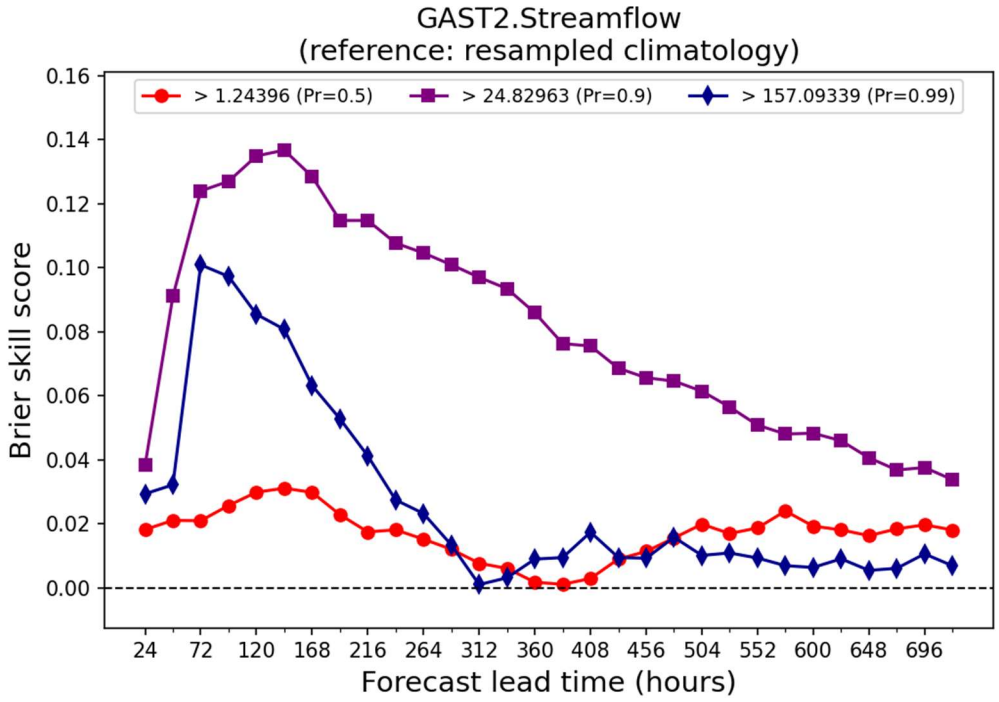

Figure 4-28: BSS of HEFS ensemble daily streamflow forecasts versus lead time at GAST2. The skill score is computed at 10, 90 and 99% quantile thresholds using ensemble streamflow forecasts forced by GEFS-climatology (control) and resampled climatology (reference).

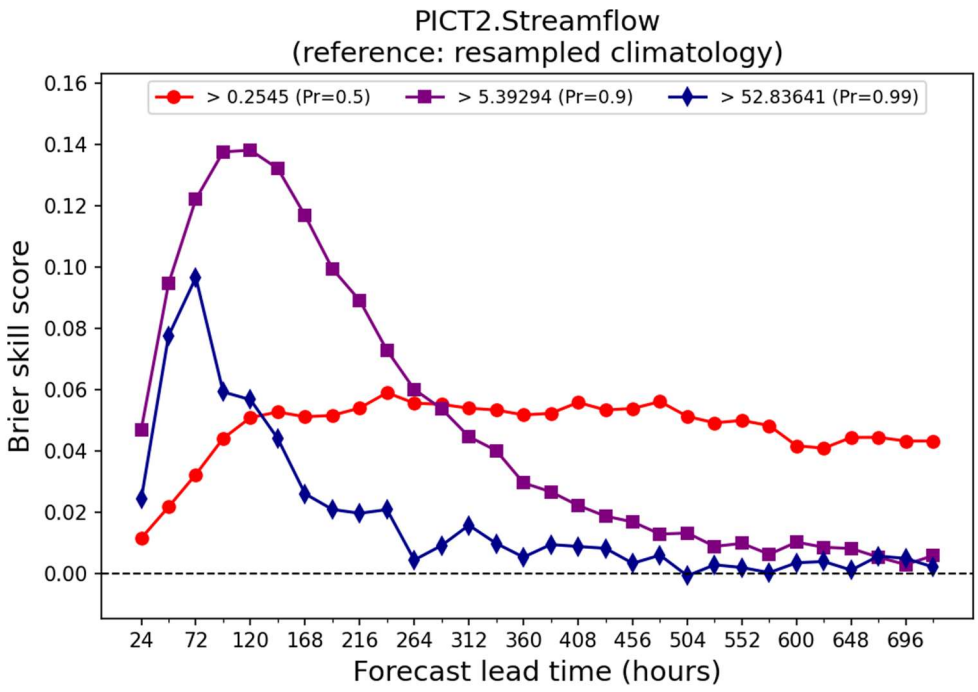

Figure 4-29: As Fig. 4-28, except at PICT2.

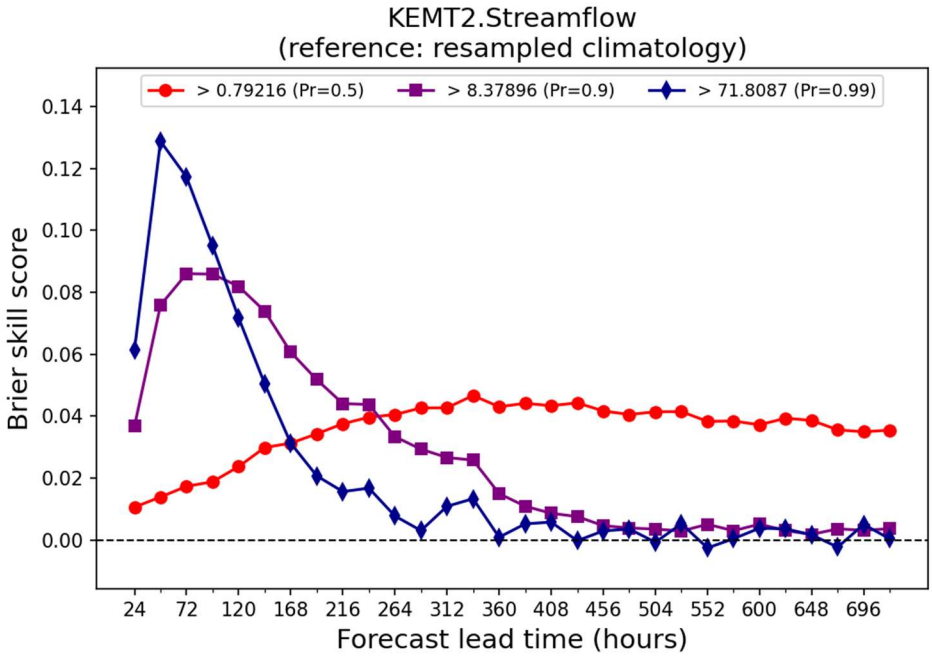

Figure 4-30: As Fig. 4-28, except at KEMT2.

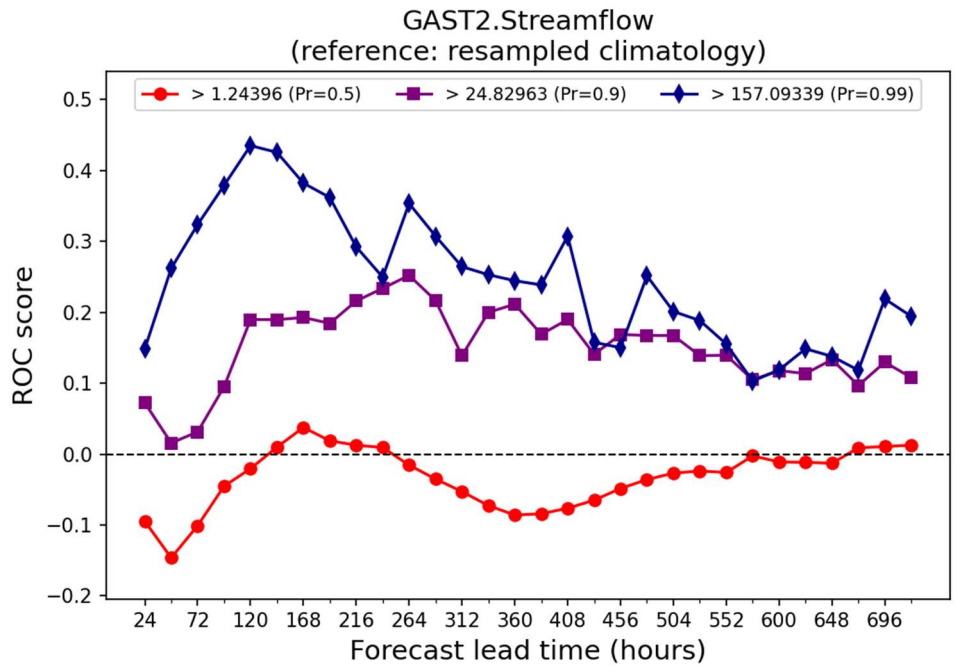

Figure 4-31: ROC score of HEFS ensemble daily streamflow forecasts versus lead time at GAST2. The skill score is computed at 10, 90 and 99% quantile thresholds using ensemble streamflow forecasts forced by GEFS-climatology (control) and resampled climatology (reference).

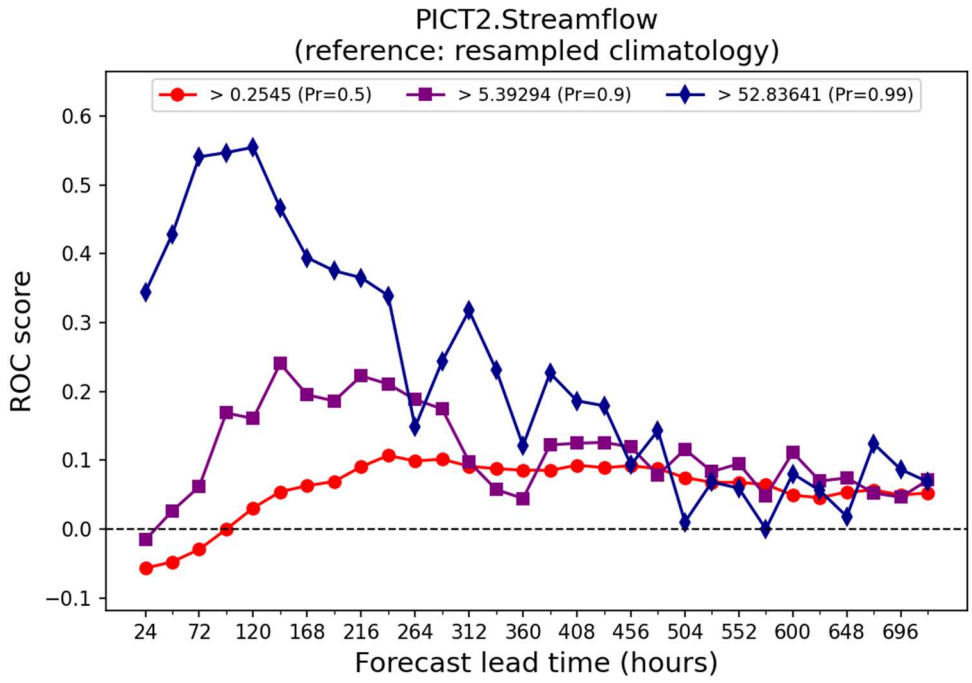

Figure 4-32: As Fig. 4-31, except at PICT2.

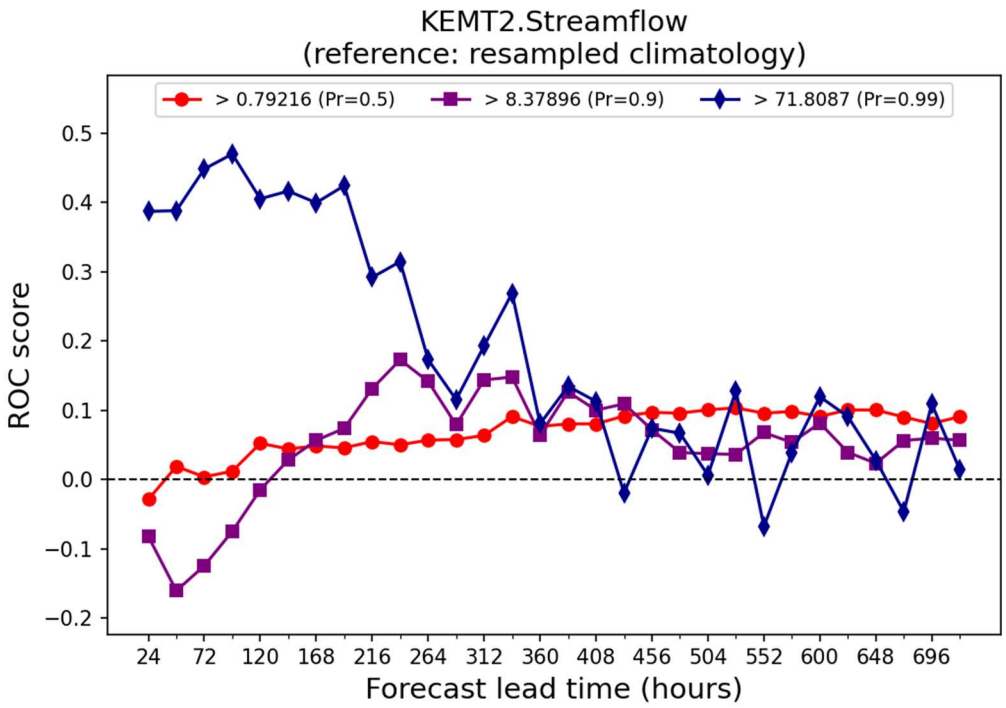

Figure 4-33: As Fig. 4-32, except at KEMT2.

Case Study: June 2007 Flood Episode

We illustrate the performance of HEFS ensemble precipitation and streamflow forecasts during the flood episode in June of 2007, which exemplifies impactful, convection-driven flood events in Texas. In the spring of 2007, a sequence of rainfall episodes produced several high flow pulses to the reservoirs in the FIRO Pilot. In Lake Georgetown, the water encroached into the flood pool by late March. In late June, a major convective outbreak produced heavy rainfall in and near the FIRO Pilot (Fig. 4-34). The rainfall was particularly intense upstream of Lake Georgetown, with a large bullseye just to the southwest of the watershed. The resulting inflow caused the reservoir level to abruptly rise more than 30 ft in a few days. The lake level crested in early July, reaching the highest the reservoir witnessed over the past 25 years (2000-2024; Fig. 4-35).

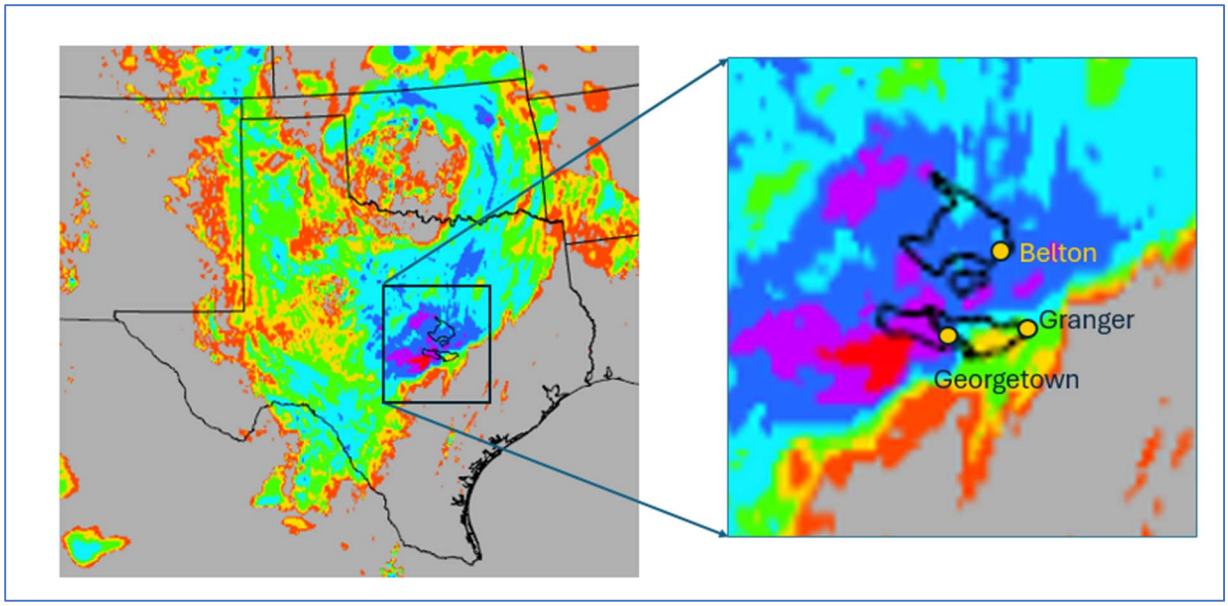

Figure 4-34: Cumulative precipitation from AORC product for the 24 hours ending on 28 June 2007.

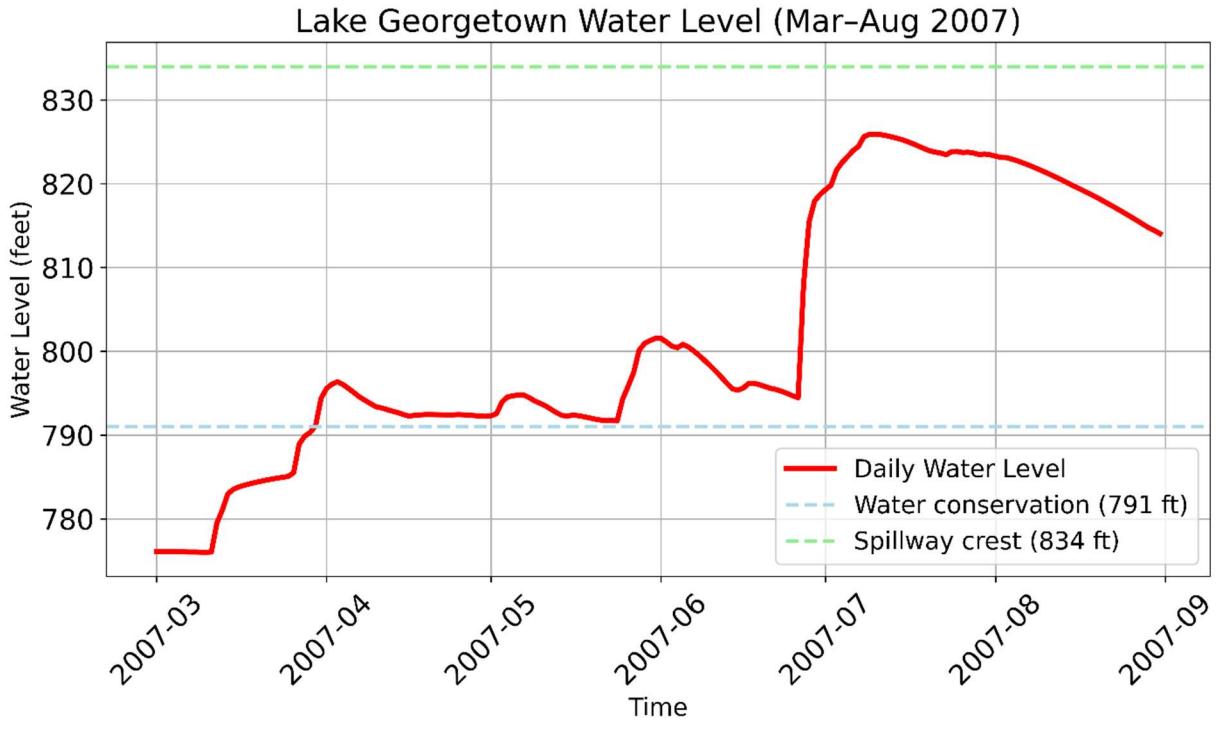

Figure 4-35: Time series of water level in Lake Georgetown from March to September 2007.

The raw precipitation forecasts from the five ensemble members of GEFSv12 reforecast data set, including one control run and four perturbed runs, are shown in Fig. 4-36. The control run forecast produces a heavy rainfall center in northern Texas that is 200 miles away from the location of observed bullseye. The rainfall distribution from perturbed runs varies widely among members: Only the second member (P2) indicates a rainfall center in central Texas, but the location is shifted to the east.

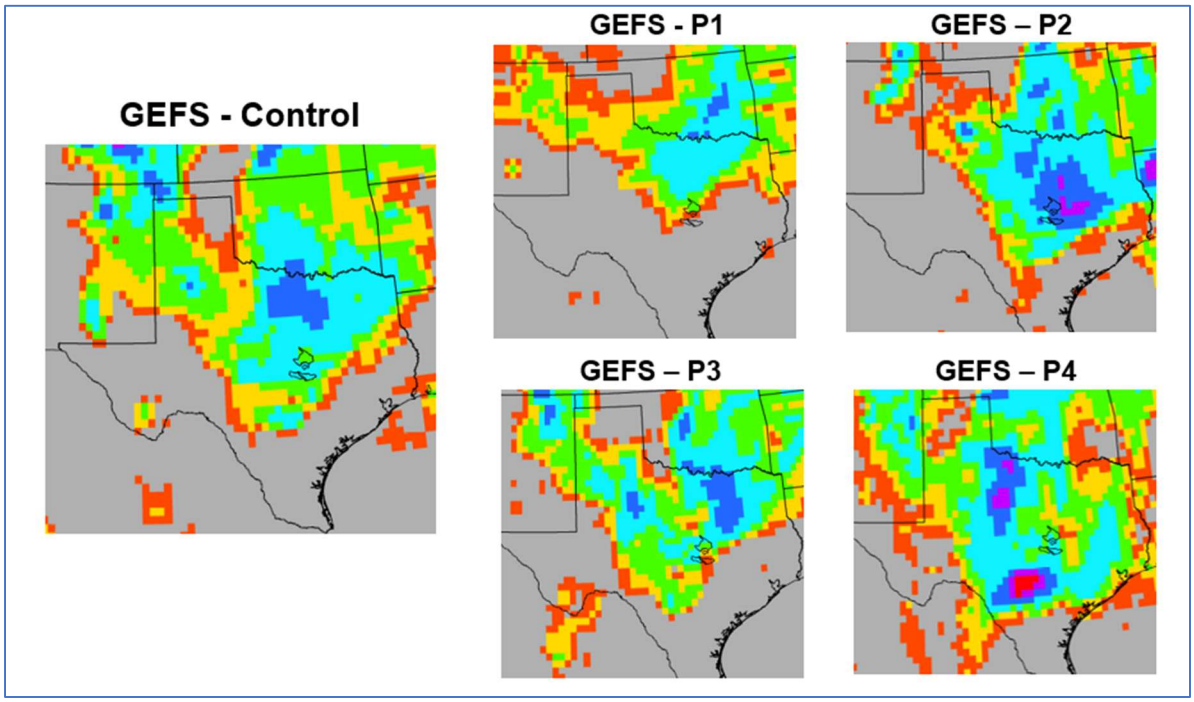

Figure 4-36: Precipitation forecasts from the ensemble members of GEFSv12 reforecast data set issued at 0z on 24 June and valid on 0z on 28 June 2007. The members include one control member (left panel), and four perturbed members (designated as P1 – P4 on the right panel).

Fig. 4-37 compares HEFS ensemble precipitation forecasts issued at 0z on 24 June 2007 against observations for the drainage upstream of Georgetown. The forecasts, comprising 40 ensemble members, were produced for the lead time range of 1–30 days using the MEFP of HEFS (Fig. 1-2). Serving as the reference is the Analysis of Record for Calibration (AORC) product from NWS. It is evident that the ensemble forecasts suffer a severe, negative bias throughout the event. At the peak of the rainfall, the observed mean areal precipitation was nearly 120mm/h, whereas the mean of ensemble mean is less than 15mm/h, and the 95% quantile is below 45 mm/h. This under-forecast is in part a result of the displacement errors in the precipitation forecasts shown in Fig. 4-36 and is likely amplified by the postprocessing algorithm in MEFP that tends to reduce the magnitude of forecast precipitation amounts.

Fig. 4-38 compares HEFS ensemble forecasts of reservoir inflows issued at 0z 24 June 2007 against reconstructed inflow by USACE. The under-forecast of inflow is even more pronounced than that of precipitation, with the 95% quantile of the forecast hardly exceeding 1000 cfs, less than 10% of the peak based on reconstructed inflow (> 18000 cfs). This under-forecast is

attributable to a combination of the under-forecast of the rainfall maxima and errors in the hydrologic model simulations.

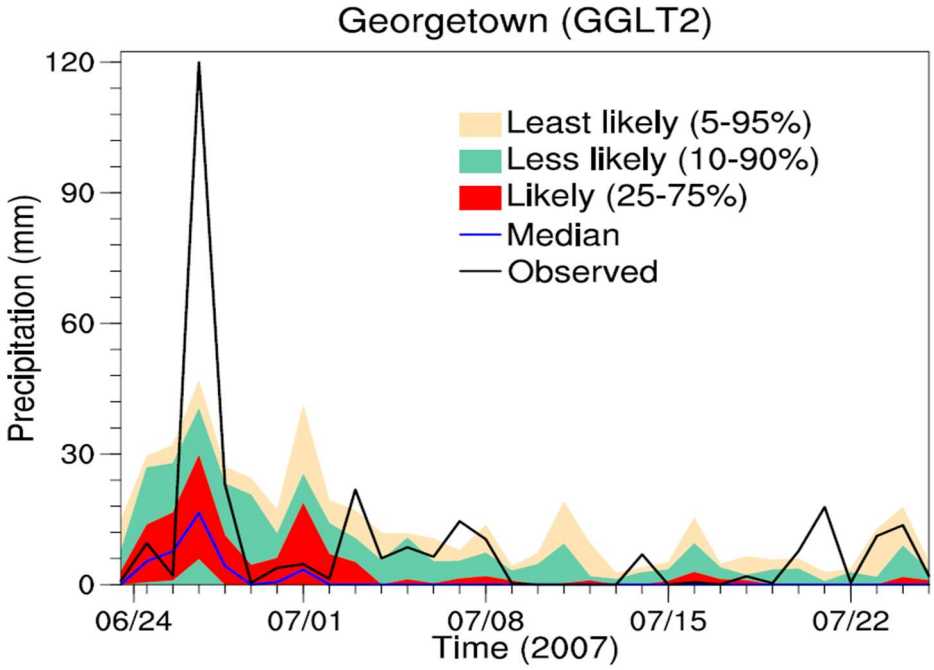

Figure 4-37: HEFS ensemble precipitation forecasts issued at 0z 24 June, 2007 for the area draining to Lake Georgetown. The forecasts were part of GEFS-Climatology suite for which postprocessed GEFSv12 reforecasts serve as forcing for days 1–14 and resampled climatology serves as forcing for day 15 and beyond.

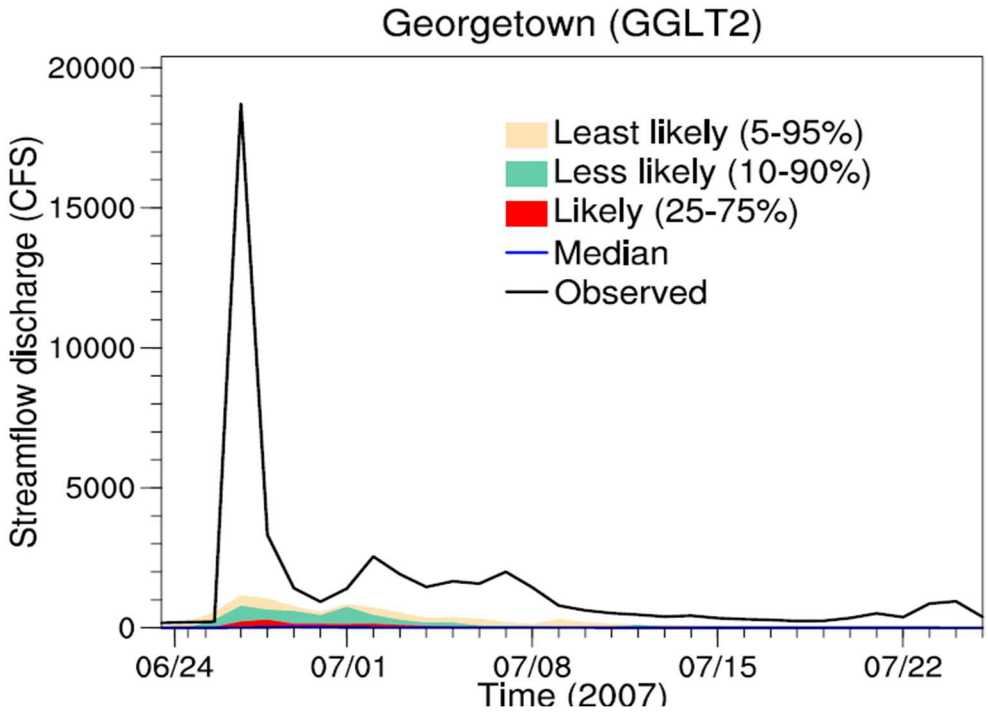

Figure 4-38: HEFS ensemble forecasts of inflow to Lake Georgetown issued at 0z on June 24, 2007. The forecasts were driven by precipitation forecast from GEFS-Climatology suite as described earlier.

Impacts of Updating SAC-SMA Parameter Values

In order to determine the impacts from updating the SAC-SMA parameter values from the 2022 calibration, we evaluate streamflow simulations performed using the two suites of parameters, namely those from the calibration completed in 2008 and 2022 (henceforth referred to as Calib-2008, and Calib-2022, respectively).

Fig. 4-39 shows summary statistics computed at each site for each suite of streamflow simulation averaged on daily time steps over the period of 2000-2019, including Percent Bias, correlation, Nash-Sutcliffe Efficiency (NSE), and Kling-Gupta Efficiency (KGE). NSE takes the following form:

$$NSE = 1 - \frac{\sum_{t=1}^{T}(Q_o(t) - Q_m(t))^2}{\sum_{t=1}^{T}(Q_o(t) - \overline{Q_o})^2}$$

Where Q_o and Q_m are observed and modeled discharge, respectively, and $\overline{Q_o}$ is the mean of observed discharge. The metric is a measure of error in the prediction versus variability (dispersion) in the observed series.

KGE is a composite measure of forecast accuracy that fuses correlation of prediction and observation, variability in prediction versus that in observation and bias. It is defined as follows:

$$KGE = 1 - ((r-1)^2 + (\alpha-1)^2 + (\beta-1)^2)^{1/2}$$

Where r is the Pearson's correlation; α is the ratio of variance of simulated and observed series; and β is bias.

Notable observations are summarized below.

- At a majority of the sites, streamflow simulations exhibit negative biases. Between the two sets of simulations, the bias tends to be worse for Calib-2022.
- The correlation from the two sets of simulations appears comparable. At a few sites (PICT2, BLNT2), correlation for the Calib-2022 suite is higher.
- At GAST2, the simulations from Calib-2008 are close to bias-neural, whereas those from Calib-2022 exhibit a slightly positive bias.
- At KEMT2 and PICT2, Calib-2022 clearly underperforms Calib-2008 for producing more severely negative biases.
- The performance of two sets of simulations varies as judged by Nash-Sutcliffe Efficiency. With Cailb-2022, a sizable improvement is seen at PICT2, whereas degradation is observed at KEMT2.

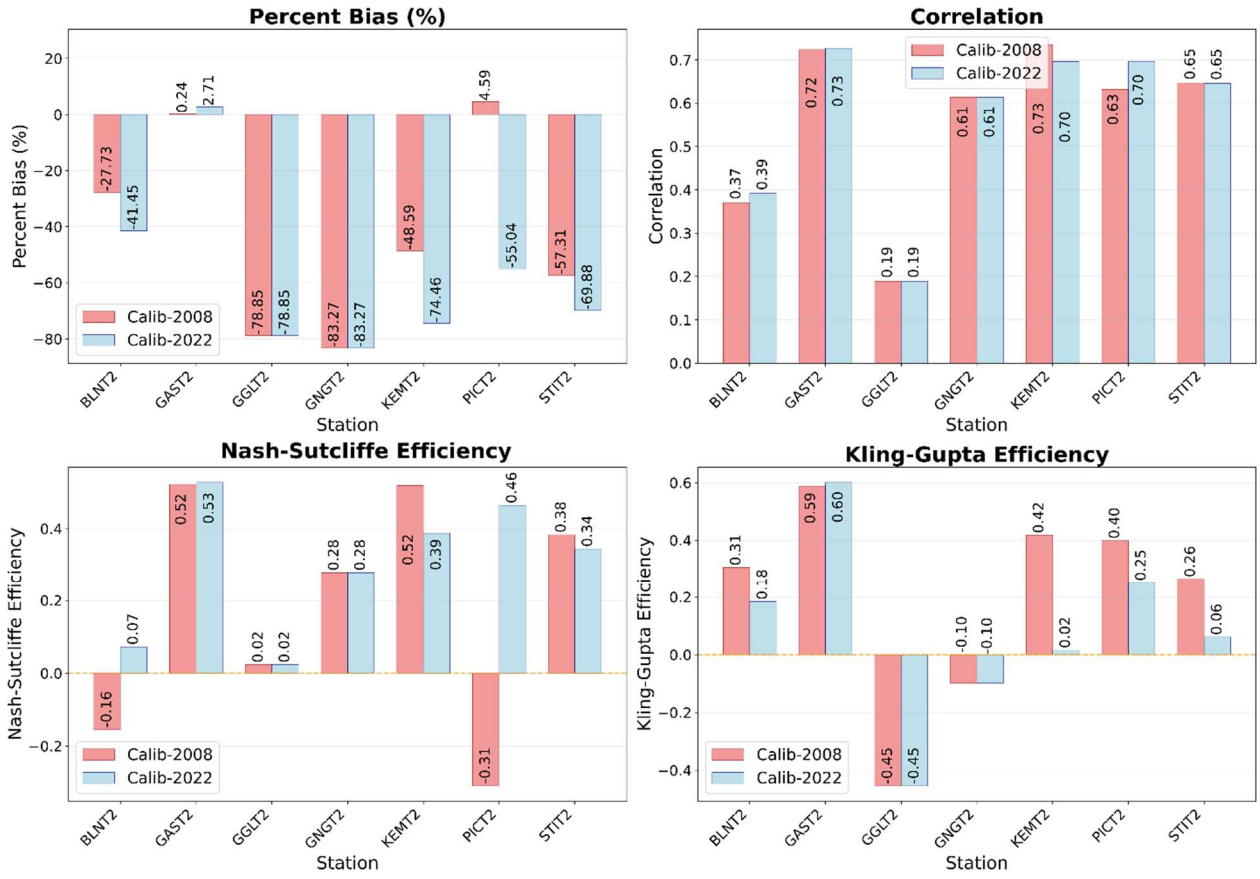

Figure 4-39: Summary statistics of streamflow simulations at each of the three

To further determin the impacts from the 2022 calibration, we plot the simulated hydrographs from Calib-2008 and Calib-2022 versus observations at the three USGS stations for the four flow episodes that feature the highest peaks. The results are shown in Figs. 4-39 – 4-42.

At GAST2, the two sets of simulations are rather similar in each of the four events (Fig. 4-40). Among these events, simulated hydrographs from both Calib-2008 and Calib-2020 severely underrepresent the magntiude of flood peak for the first three (May 1990, December, 1992, and June 2007), and the simulations using the earlier parameter values (Calib-2008) perform slightly better by featuring high peak values. For the October 2018 event, both simulated hydrographs closely reproduce the observed one, with that from Calib-2022 performining better in terms of peak mangitude and timing.

At PICT2, the differences between the two sets of simulations are pronounced for the four flooding events (Fig. 4-41). Between Calib-2008 and Calib-2022, the simulations from the latter consistently underperform by producing much smaller peak discharge for each event. Those from Calib-2008 fare better, though still tend to biased low for three out of four events.
At KEMT2, the differences between the two sets of simulations are also quite pronounced for the four flooding events (Fig. 4-42). As in the case of PICT2, simulations from Calib-2008 feature much higher peak discharge than those from Calib-2022 across all four events, whereas latter are biased low consistently. Note that for the first two events, which occurred in December 1991 and March 1998, simulated peaks from Calib-2008 overshoot and are positively biased.

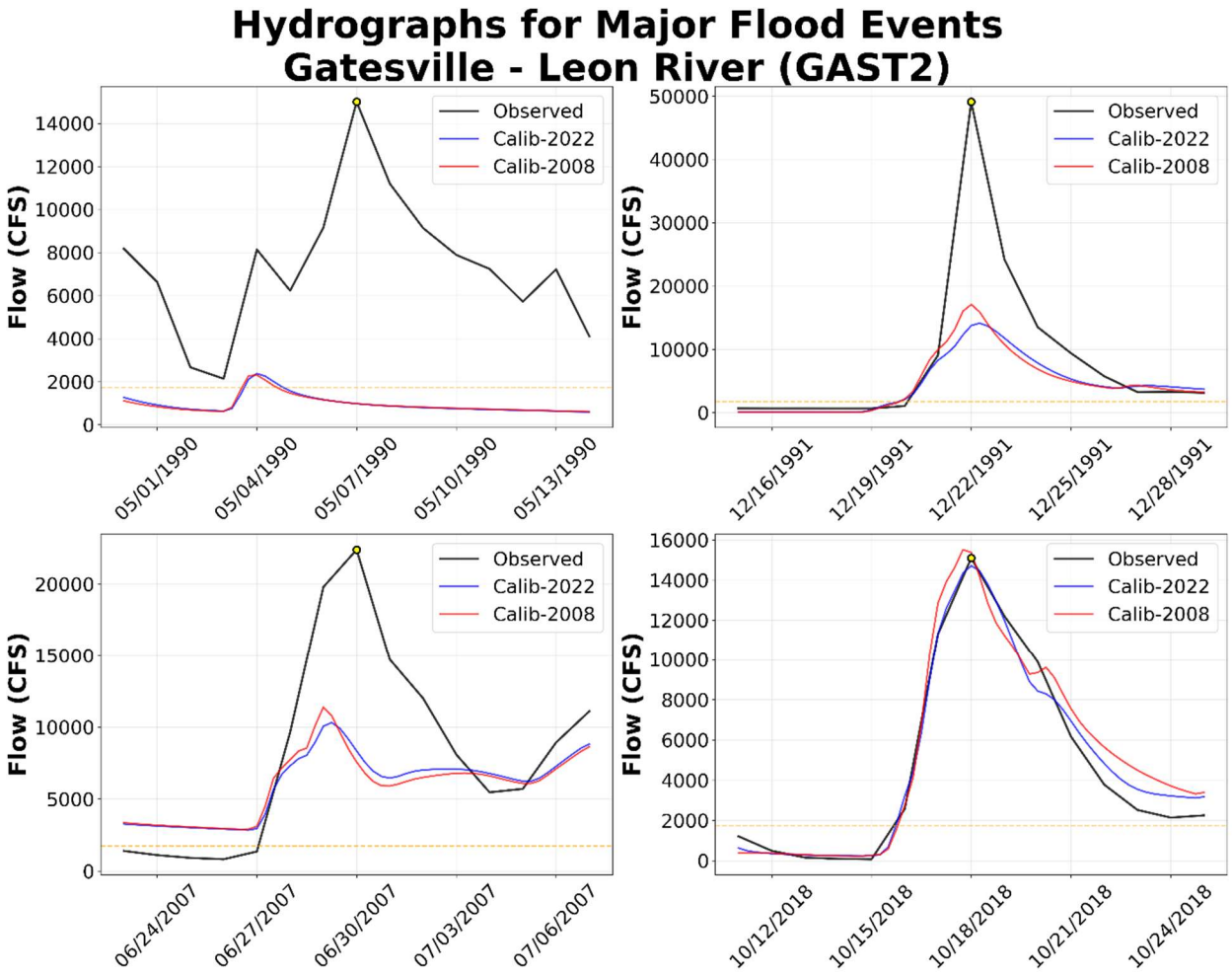

Figure 4-40: Simulated and observed hydrographs for four events with the largest peak flow over 1990-2019 at GAST2, which occurred in: 1) May 1990; 2) December, 1991; 3) June 2007, and 4) October 2018.

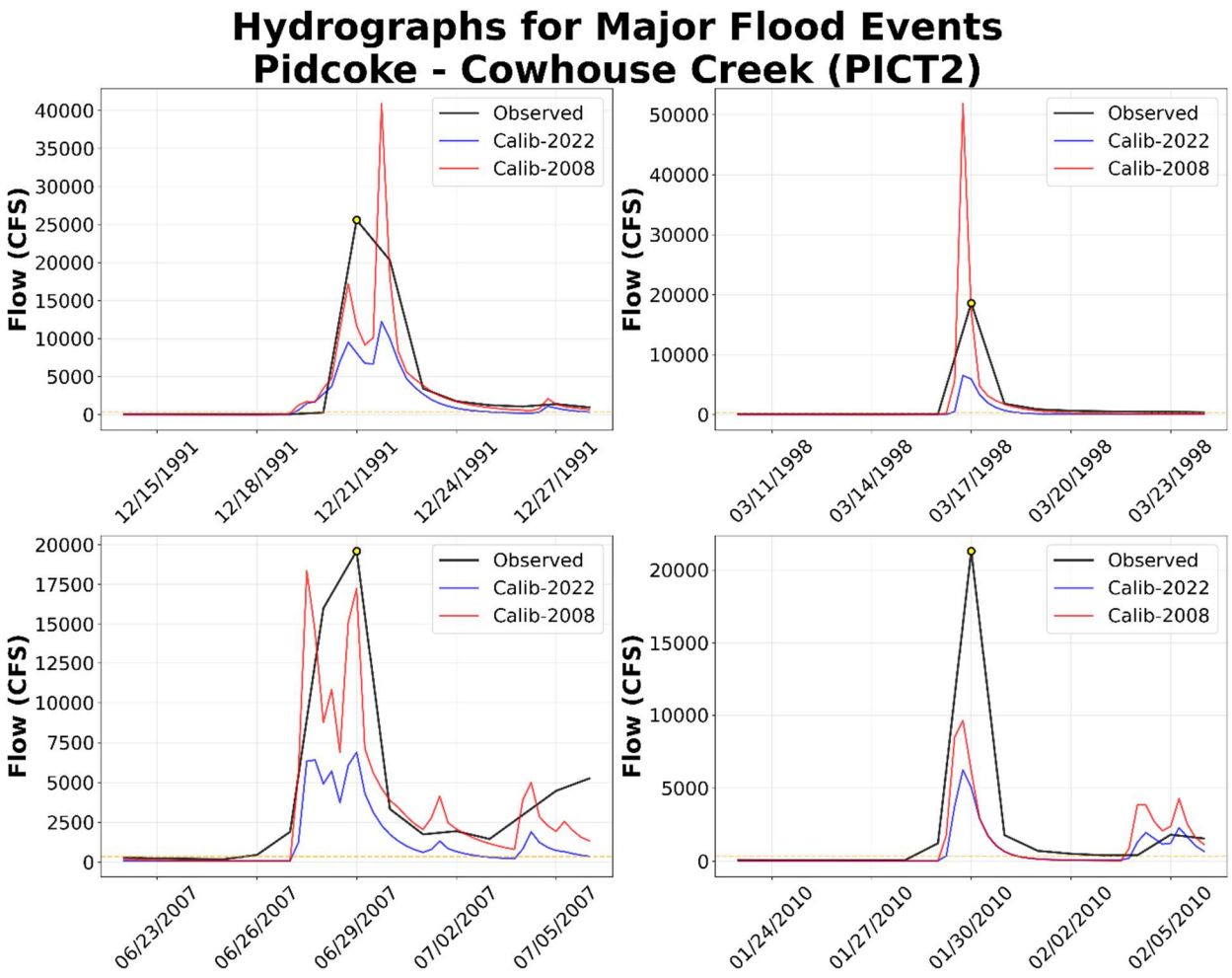

Figure 4-41: Simulated and observed hydrographs for four events with the largest peak flow over 1990-2019 at PICT2, which occurred in: 1) December 1991; 2) March 1998; 3) June 2007, and 4) January 2010.

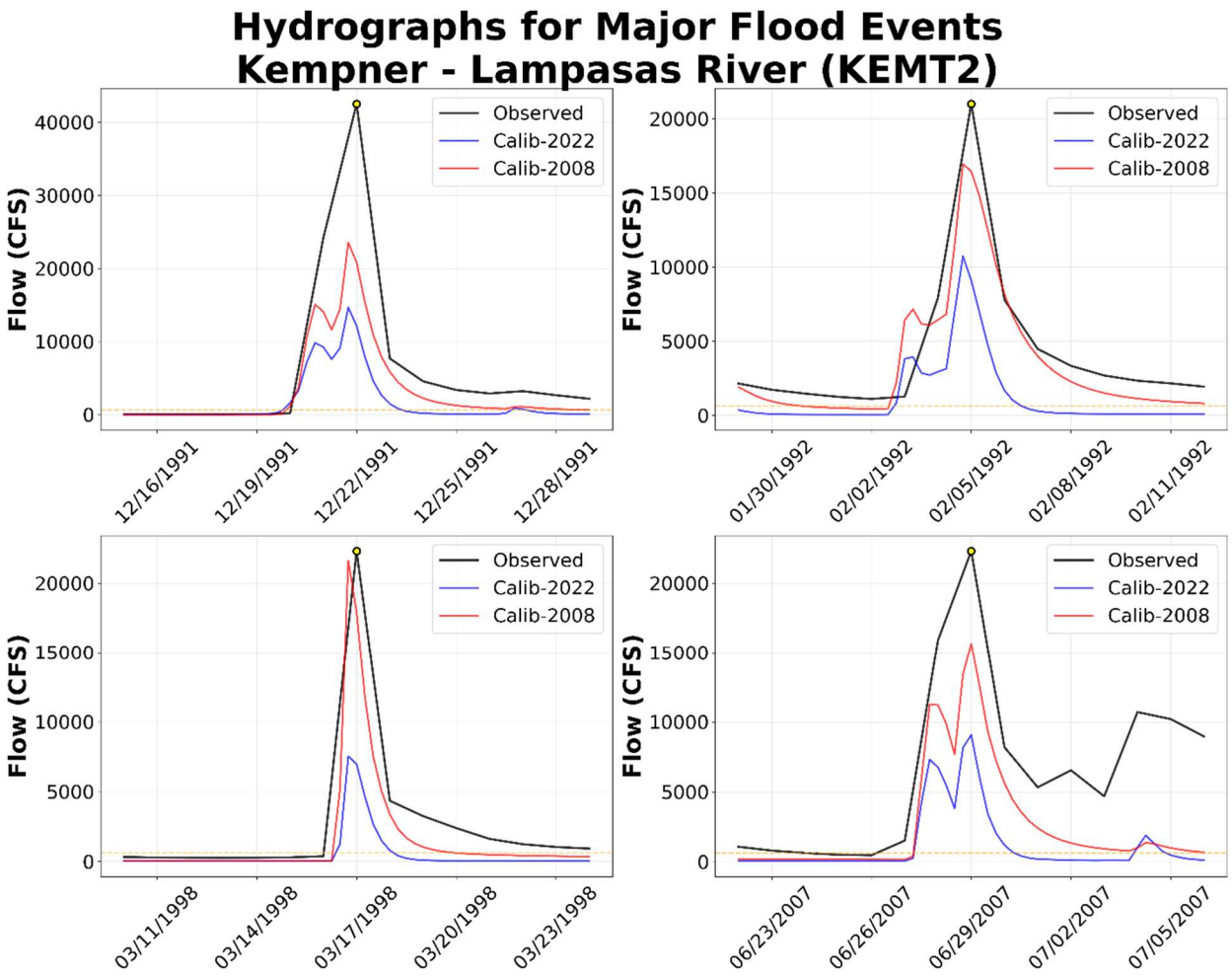

Figure 4-42: Simulated and observed hydrographs for four events with the largest peak flow over 1990-2019 at KEMT2, which occurred in: 1) December 1991; 2) February 1992; 3) March 1998, and 4) June 2007.

4.2. GEFS-S2S:

The GEFSv12 S2S forecasts are ingested into HEFS to produce extended-range streamflow forecasts that undergo evaluations. To simplify matters, we will focus on BSS and ROC scores for precipitation and streamflow forecasts at the three forecast points collocated with USGS stations, namely GAST2, PICT2 and KEMT2.

The BSS and ROC scores for HEFS ensemble precipitation forecasts are shown in Figures 4-43 – 4-48. Notable observations are summarized below.

- At all three sites, BSS declines with lead time and threshold. At the middle and top thresholds, the skill remains slightly positive until weeks 3–4, but it is evident that most of the skill is contained in the week 1 forecasts.
- ROC scores indicate that the discrimination skills of HEFS precipitation forecast are the highest at the top threshold (90%9), and lowest at the bottom threshold (50%). The scores of the former threshold are mostly positive across the lead time range (through week 8).
- At the middle (90%) and bottom (50%) thresholds, ROC scores diminish beyond week 2.

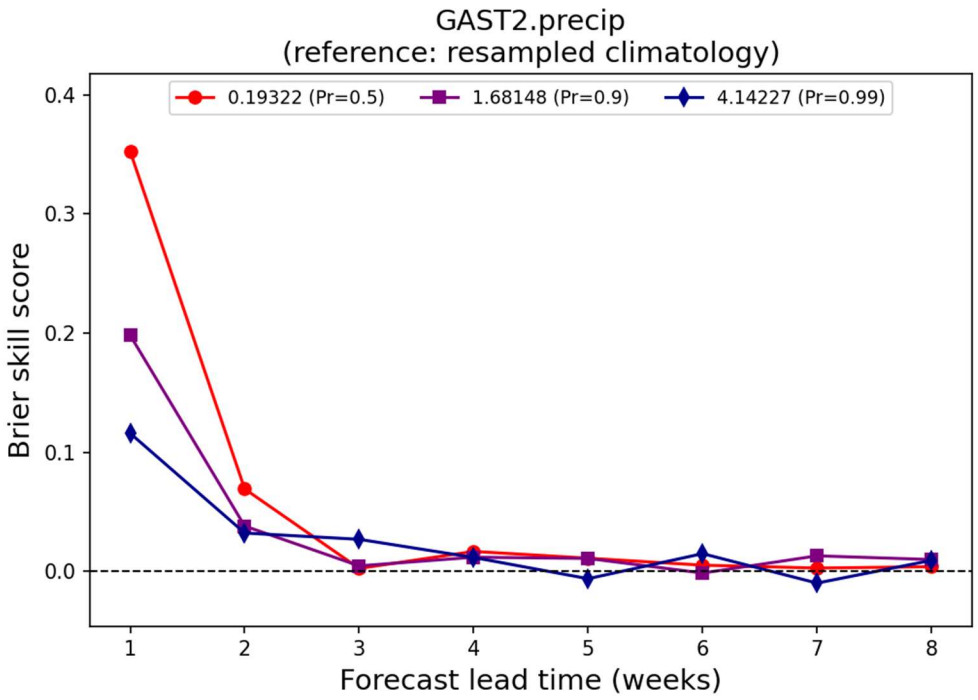

Figure 4-43: BSS of HEFS ensemble precipitation forecasts against lead time at GAST2. The skill score is computed at 50, 90 and 99% quantile thresholds on postprocessed GEFSv12 S2S forecasts aggregated onto weekly intervals, with the climatological probabilities serving as the reference.

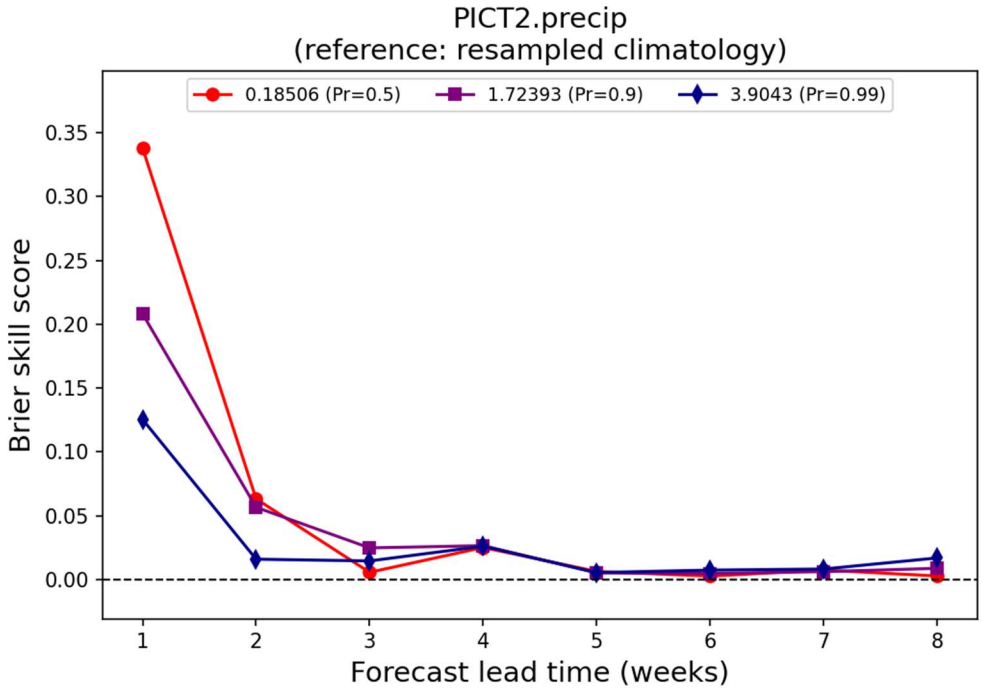

Figure 4-44: As Fig. 4-38, except at PICT2.

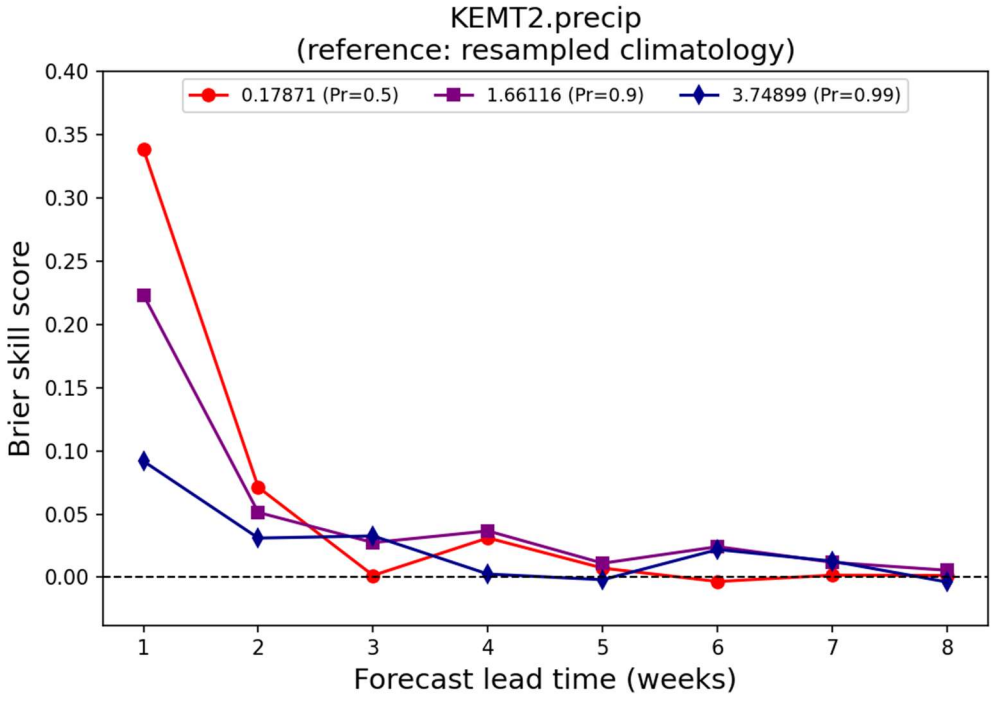

Figure 4-45: As Fig. 4-38, except for KEMT2.

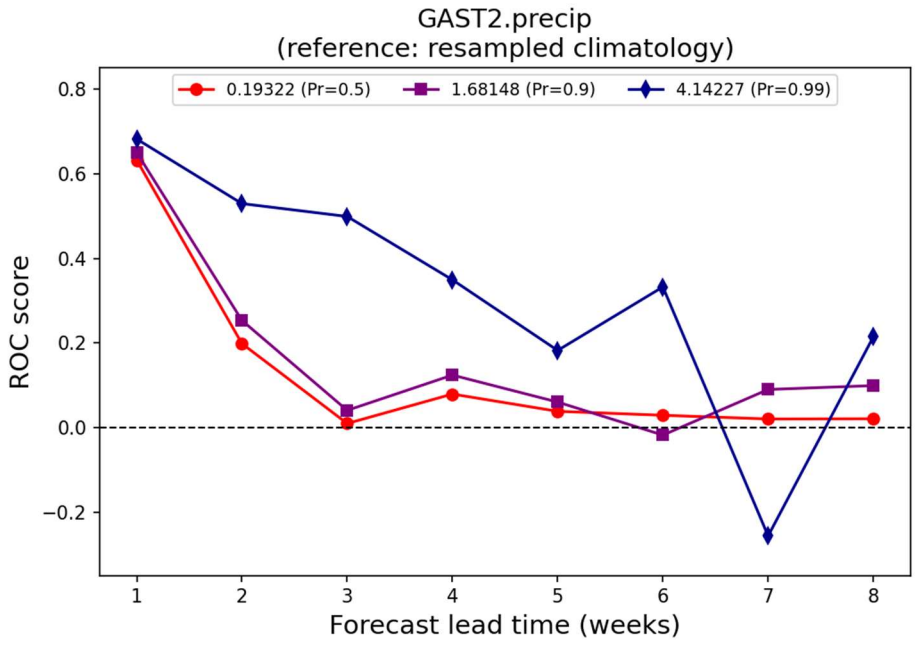

Figure 4-46: ROC scores of HEFS ensemble precipitation forecasts against lead time at GAST2. The skill score is computed at 50, 90 and 99% quantile thresholds on postprocessed GEFSv12 S2S forecasts aggregated onto weekly intervals, with the climatological probabilities serving as the reference.

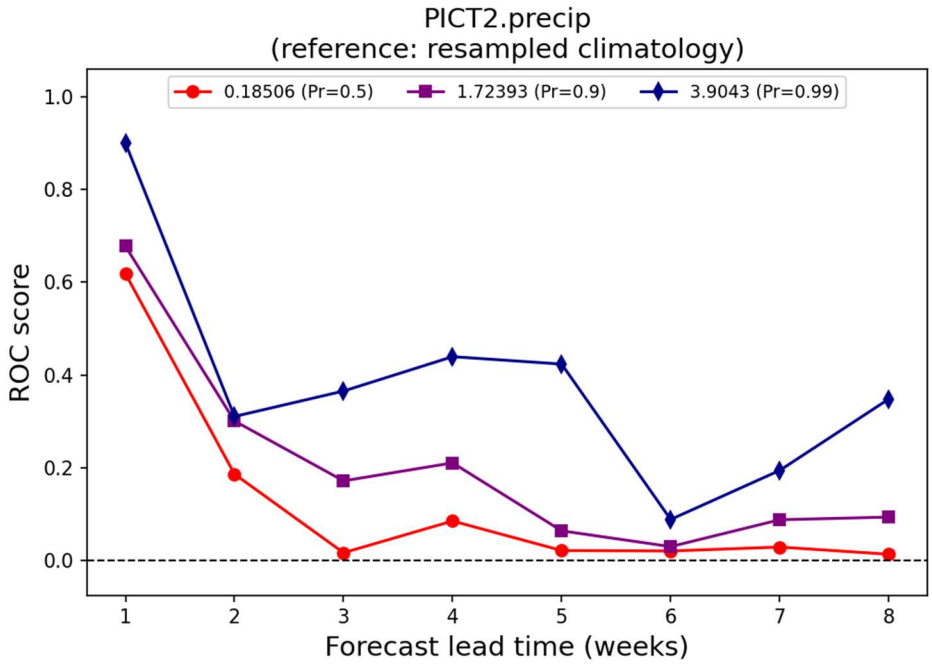

Figure 4-47: As Fig. 4-41, except at PICT2.

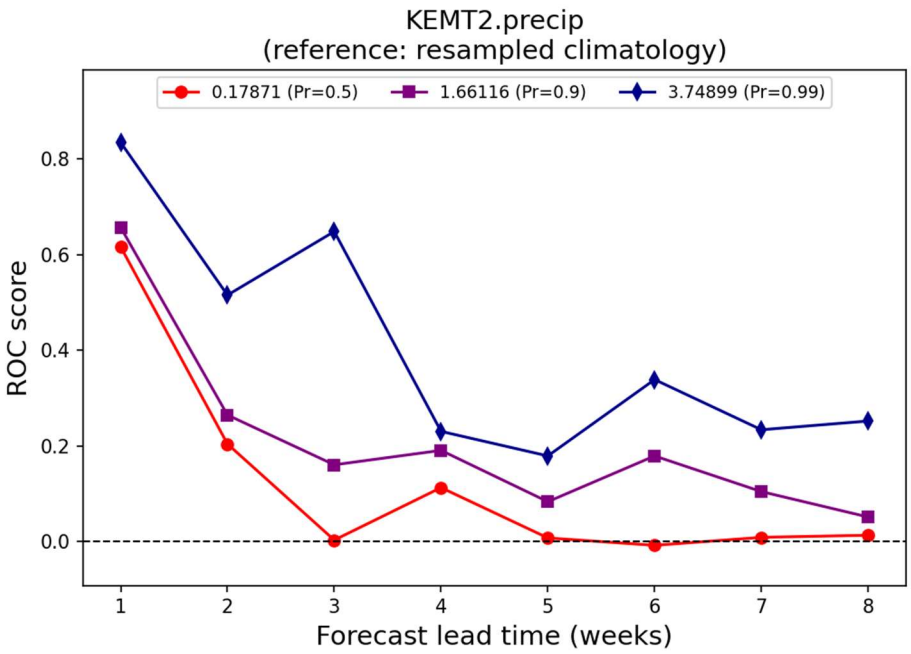

Figure 4-48: As Fig. 4-41, except at KEMT2.

The BSS and ROC scores for HEFS ensemble streamflow forecasts are shown in Figures 4-49 – 4-54. Notable observations are summarized below.

- As judged by BSS, the skills of ensemble streamflow forecasts vary greatly among the three sites and thresholds.
    - At the middle threshold (90%), the forecasts are consistently skillful at all three sites out to week 8, though the skills tend to decline with lead time.
    - At the top threshold (90%), the skills are broadly lower at each site than those at the middle threshold. The BSS tends to decline with lead time at GAST2 and KEMT, but no clear trend is observed at PICT2.
    - At the bottom threshold (10%), The BSS is flat at GAST2, suggesting no skills relative to climatology. At two other sites, BSS exhibits a rising trend with lead time.
- ROC scores at the three sites vary widely.
    - At GAST2 and PICT2, it appears that that the discrimination skills of HEFS precipitation forecast are the highest at the top threshold (90%9), and lowest at the bottom threshold (10%), whereas at KEMT2, ROC scores at the top threshold are broadly lower than those at the middle threshold.
    - The dependence of ROC scores on lead time varies among sites and thresholds. At the lowest threshold (10%), the ROC scores do not exhibit a clear downward trend.
    - At the highest threshold (90%), the ROC scores tend to be lower at PICT2 and KEMT2, both associated with small drainage areas.

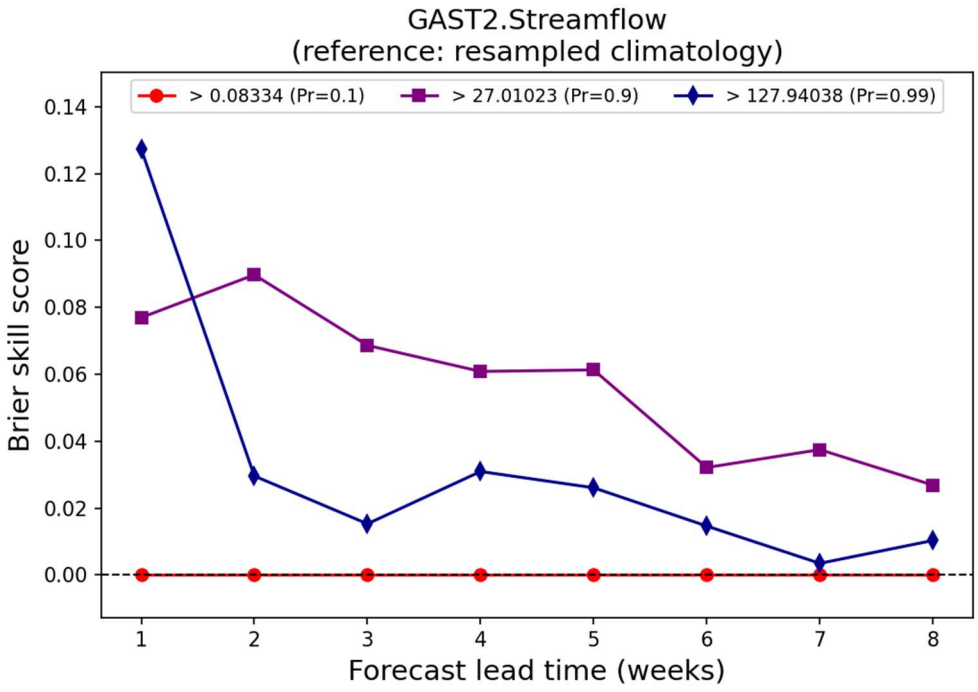

Figure 4-49: BSS of HEFS ensemble streamflow forecasts against lead time at GAST2. The skill score is computed at 10, 90 and 99% quantile thresholds on streamflow forecasts forced by postprocessed GEFSv12 S2S precipitation forecasts, with the ensemble streamflow forecasts driven by resampled precipitation serving as the reference. Note the streamflow forecasts are aggregated onto weekly intervals.

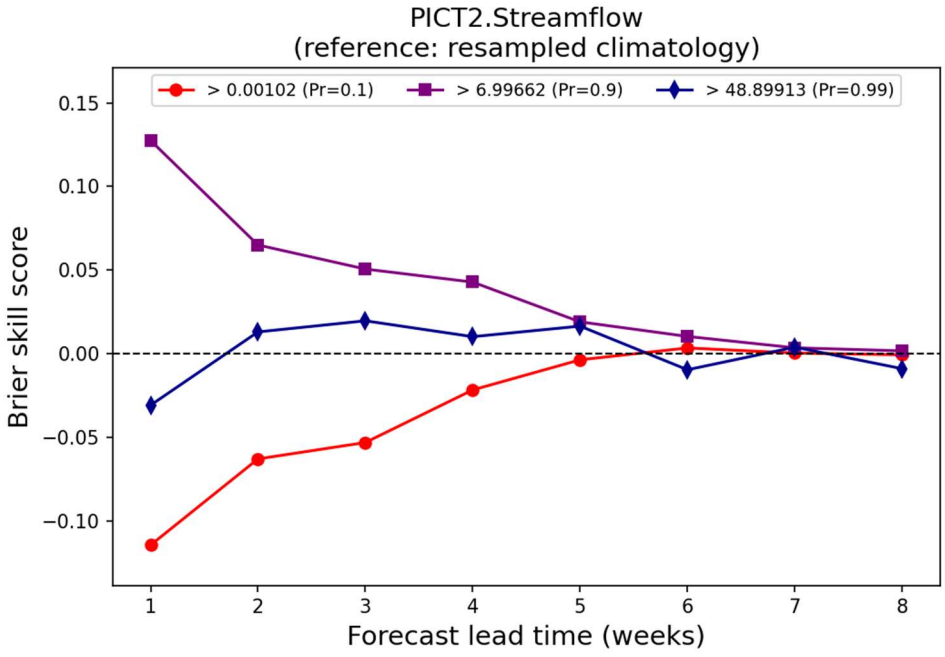

Figure 4-50: As Fig. 4-49 but at PICT2.

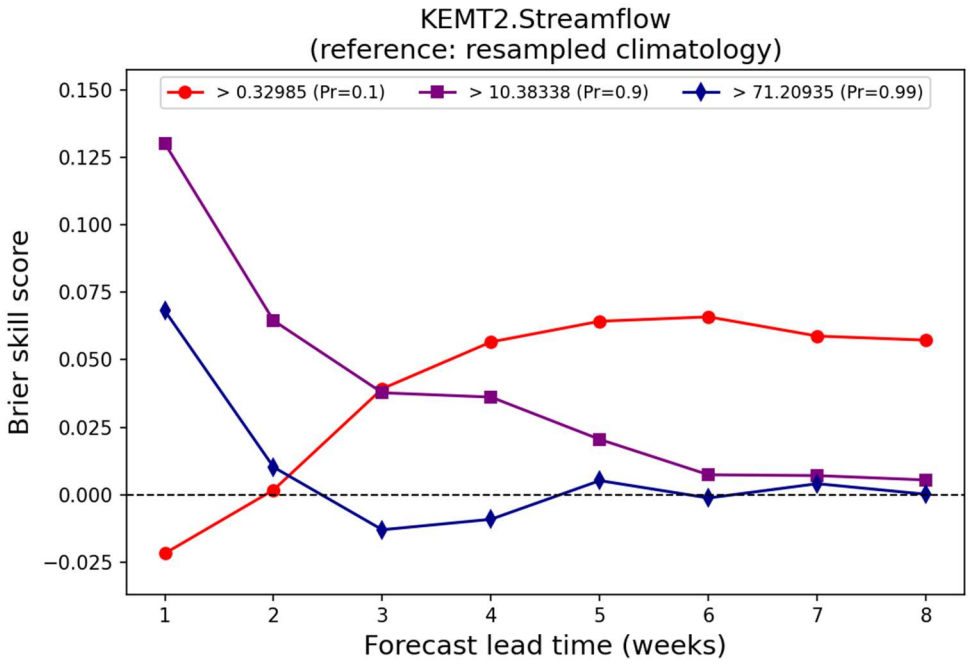

Figure 4-51: As Fig. 4-49 but at KEMT2.

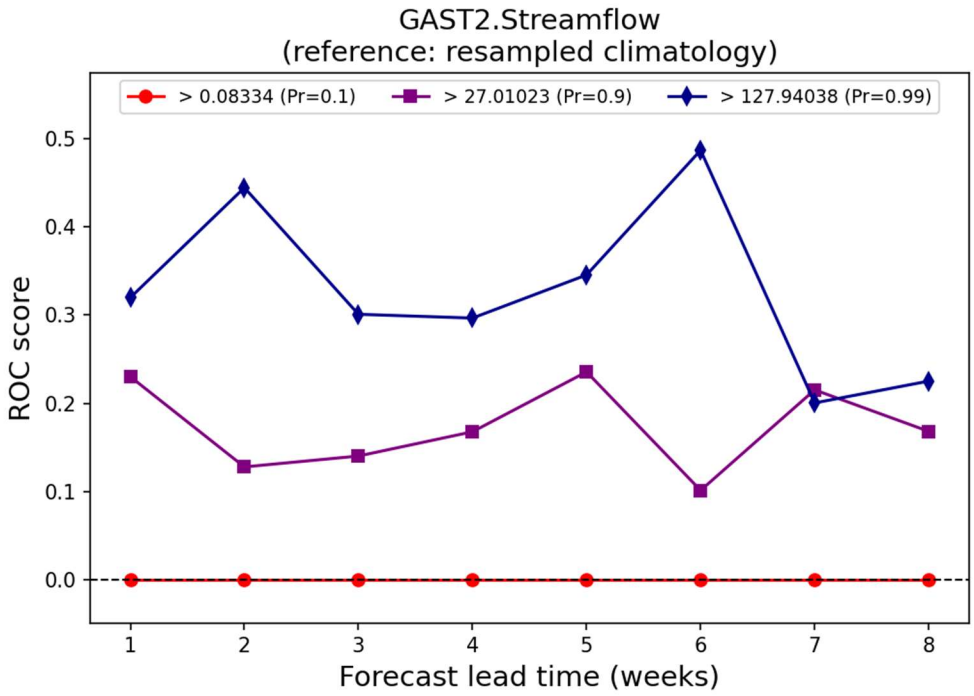

Figure 4-52: ROC score of weekly streamflow above thresholds of 10, 90 and 99% quantiles at GAST2.

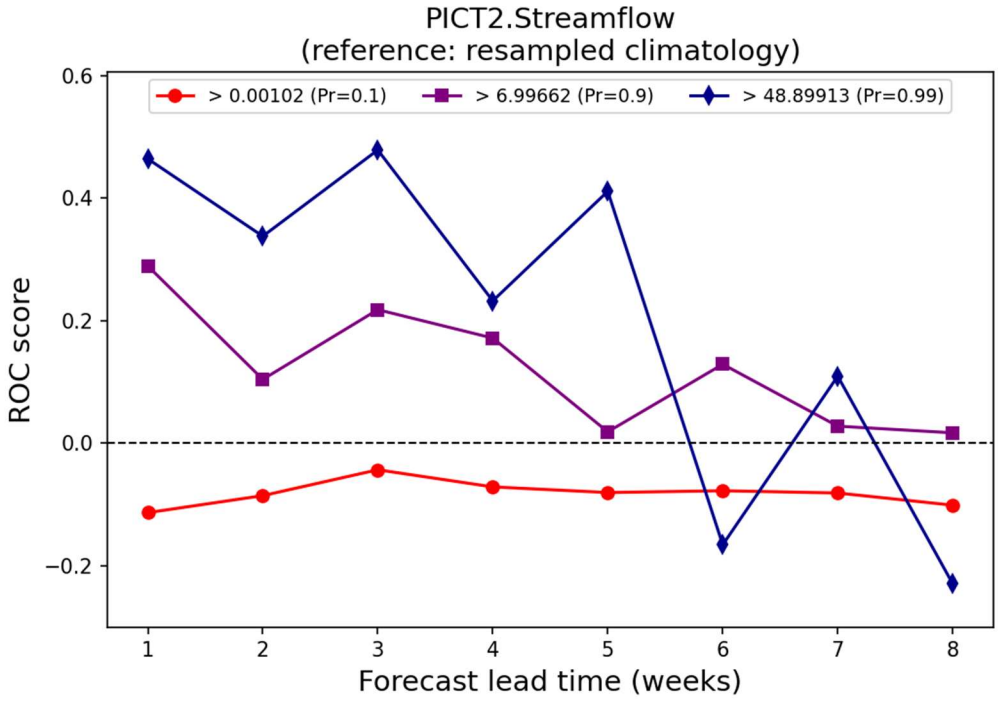

Figure 4-53: As Fig. 4-52, except at PICT2.

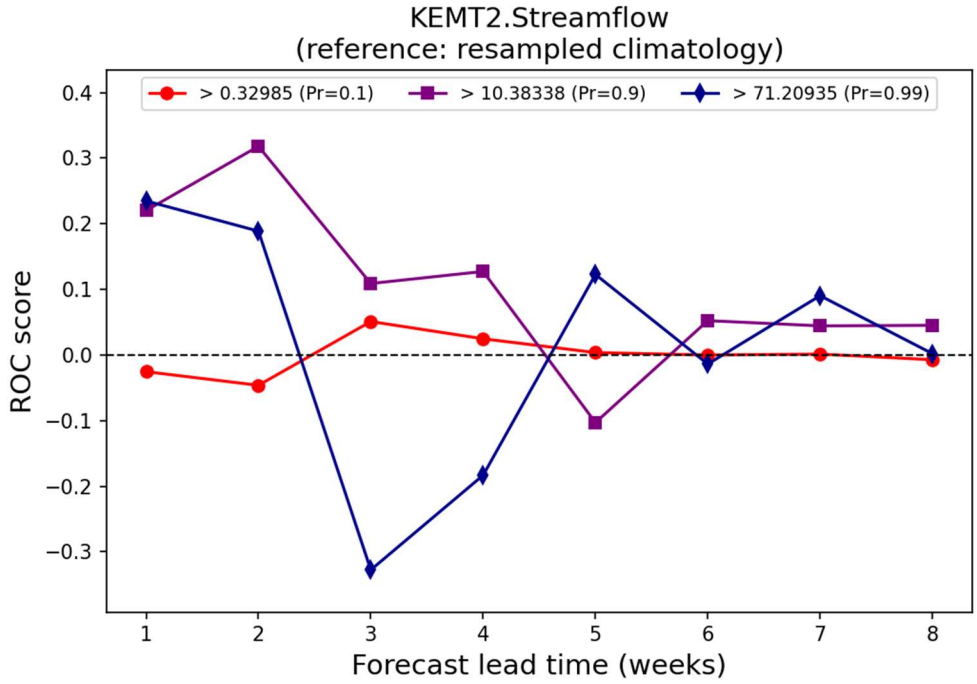

Figure 4-54: As Fig. 4-52, except at KEMT2.

We perform limited comparisons between the streamflow forecasts produced by GEFS-S2S vs. GEFS-Climatology to determine potential gains in forecast skills by replacing resampled climatology with S2S precipitation forecasts beyond day 14. Figs. 4-55 – 4-60 show the BSS and ROC scores computed for streamflow forecasts forced by GEFS-S2S with those driven by GEFS-Climatology as reference over a lead time range of 1–4 weeks. Key observations are summarized as follows:

- The results are rather mixed. At GAST2, the streamflow forecasts from GEFS-S2S appear more skillful in terms of BSS at the top threshold throughout the lead time range, but at the two lower thresholds the BSS is close to zero, suggesting marginal skills in comparison to the reference forecast.
- At PICT2 and KEMT2, BSS at the top threshold is consistently negative, whereas it remains close to zero at the other two thresholds. There appears to be a tendency for the BSS to rise with lead time, suggesting positive impacts from ingesting the GEFS S2S precipitation forecasts.
- In terms of ROC scores, there is no clear indication that the streamflow forecasts from GEFS-S2S outperforms the reference.

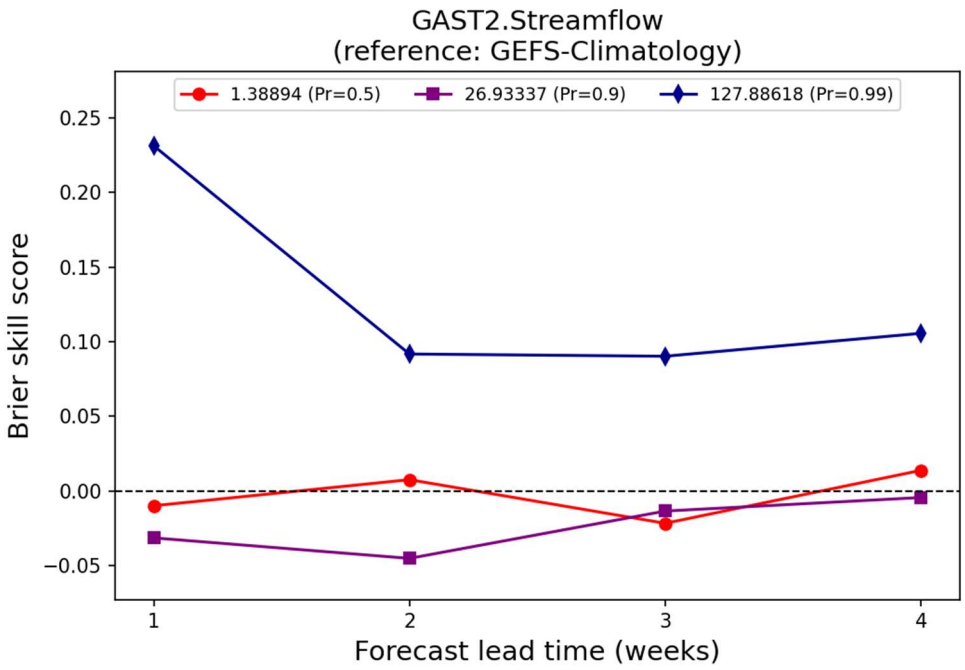

Figure 4-55: BSS of HEFS ensemble streamflow forecasts against lead time at GAST2. The skill score is computed at 10, 90 and 99% quantile thresholds on streamflow forecasts forced by postprocessed GEFSv12 S2S precipitation forecasts, with the ensemble streamflow forecasts driven by GEFS-Climatology serving as the reference. Note the streamflow forecasts are aggregated onto weekly intervals.

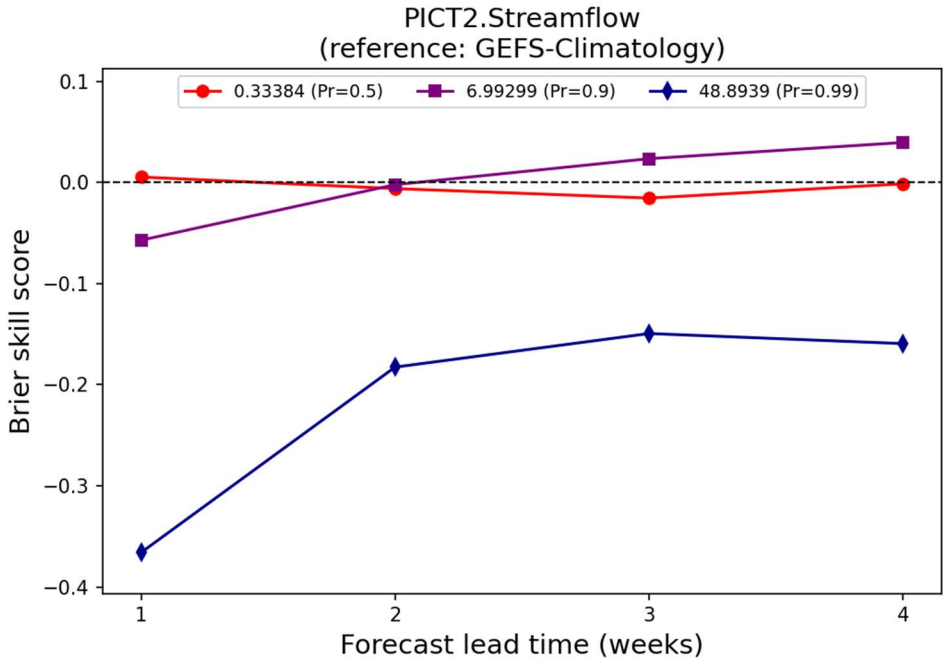

Figure 4-56: As Fig. 4-55, except at PICT2.

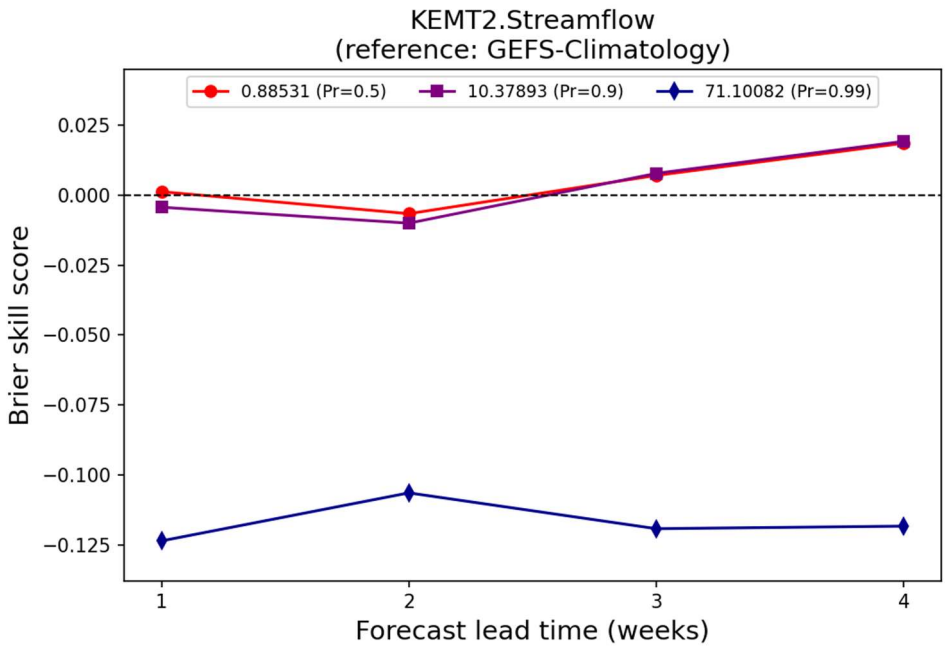

Figure 4-57: As Fig. 4-55, except at KEMT2.

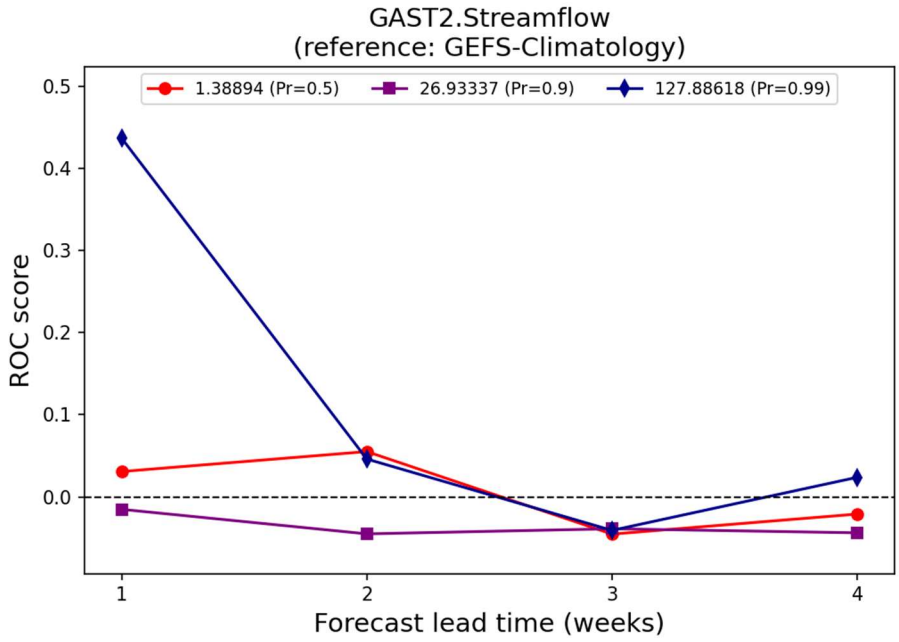

Figure 4-58: ROC score computed on HEFS ensemble streamflow forecasts against lead time at GAST2. The score is computed at 10, 90 and 99% quantile thresholds on streamflow forecasts forced by postprocessed GEFSv12 S2S precipitation forecasts, with the ensemble streamflow forecasts driven by GEFS-Climatology serving as the reference. Note the streamflow forecasts are aggregated onto weekly intervals.

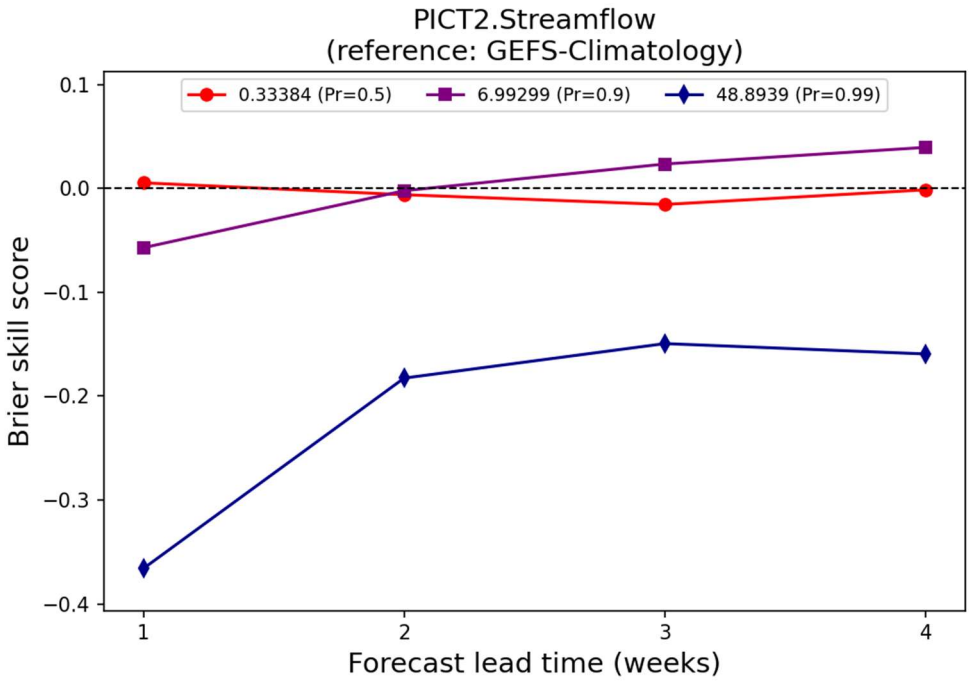

Figure 4-59: As Fig. 4-58, except at PICT2.

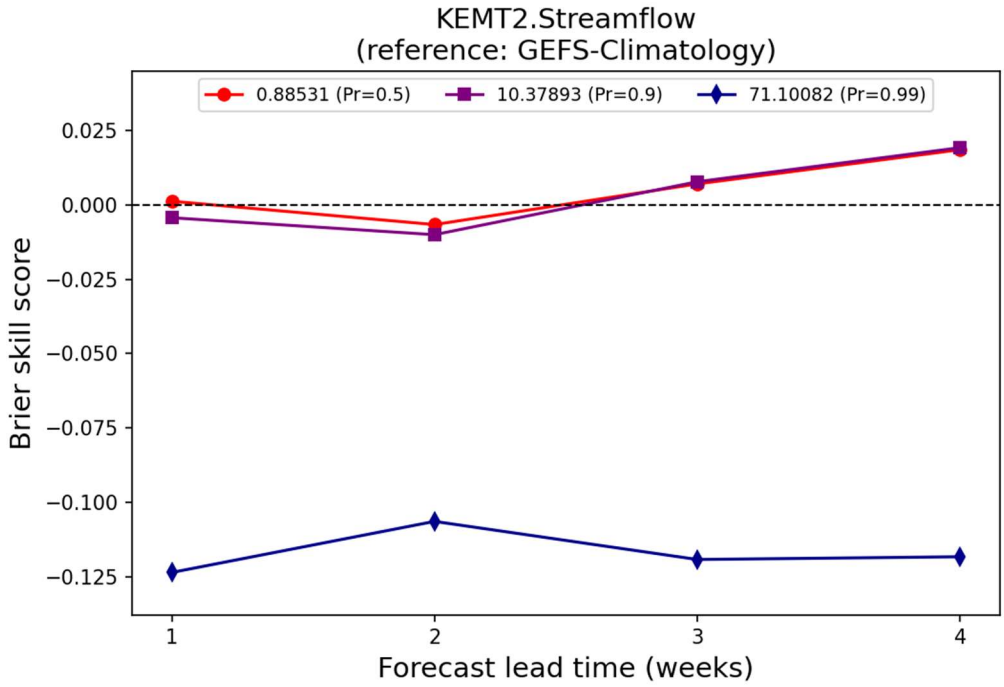

Figure 4-60: AS Fig. 4-58, except at KEMT2.

4.3. GEFS-CFSv2:

The verification of HEFS ensemble precipitation and streamflow forecasts again focuses on the three forecast points collocated with USGS stations, and only the BSS and ROC scores are shown here. In this case, the reference forecasts in computing the skill scores are HEFS precipitation and streamflow forecasts from GEFS-Climatology. The skill scores help determine the potential benefits of replacing the resampled climatology with CFSv2 forecasts for day 14 – 270.

The BSS and ROC scores thus computed are shown in Figs. 4-61 – 4.66. Key observations include the following:

- Skills in the CFSv2 precipitation forecasts are marginal against resampled climatology as judged by both BSS and ROC scores, irrespective of thresholds.
- At the lowest threshold (10%), there appears to be a tendency for BSS to be positive within the 1-5 months range, suggesting skills in forecasting dry spells. The skills are more pronounced at the forecast points in the south, particularly at KEMT2.
- As judged by ROC scores, the discrimination skills of GEFS-CFSv2 precipitation forecasts are broadly lower than those of GEFS-Climatology, suggesting that CFSv2 precipitation forecasts feature no skills relative to resampled climatology.

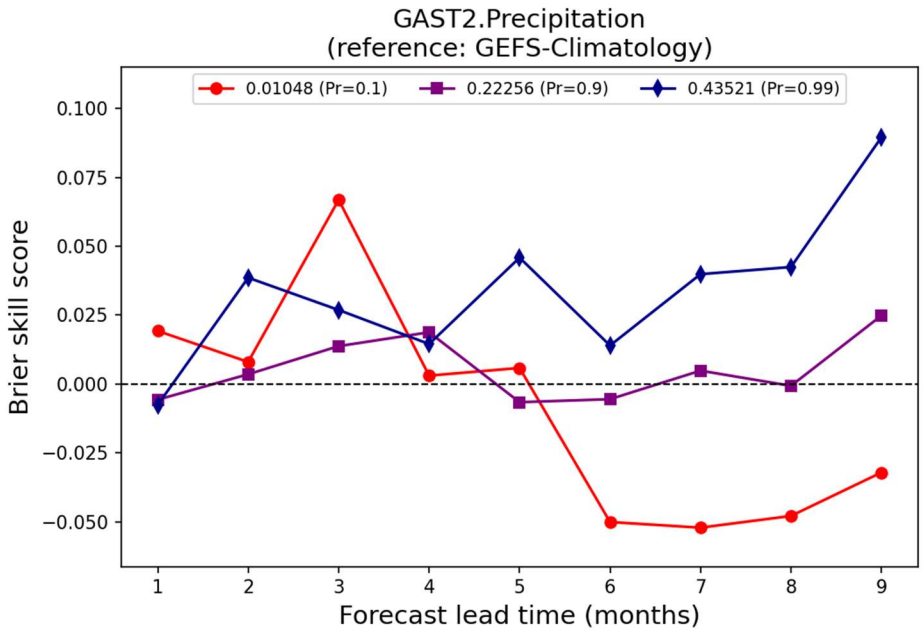

Figure 4-61: BSS of HEFS ensemble precipitation forecasts from GEFS-CFSv2against lead time at GAST2. The skill score is computed at 10, 90 and 99% quantile thresholds on postprocessed precipitation forecasts aggregated onto monthly intervals, with the climatological probabilities serving as the reference.

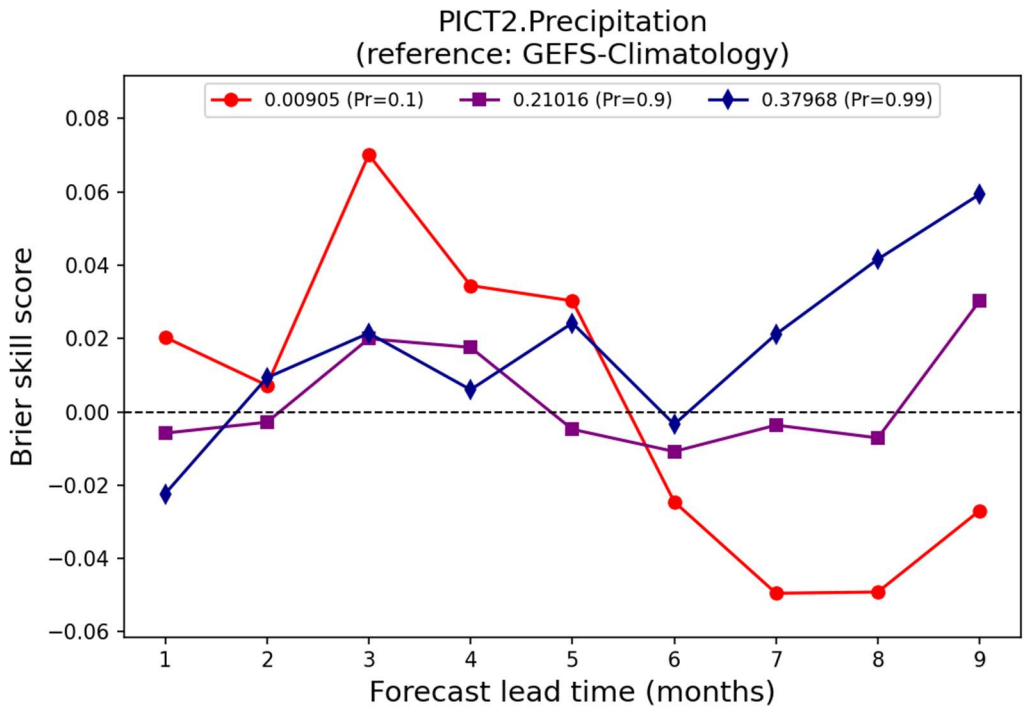

Figure 4-62: As Fig. 4-61, except at PICT2.

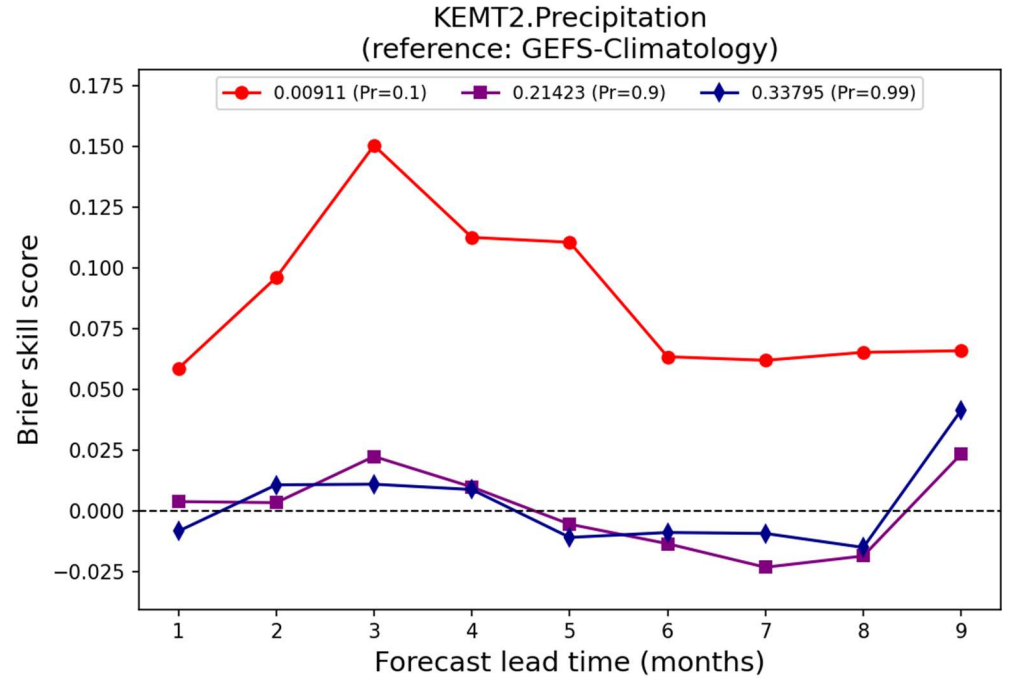

Figure 4-63: As Fig. 4-61, except at KEMT2.

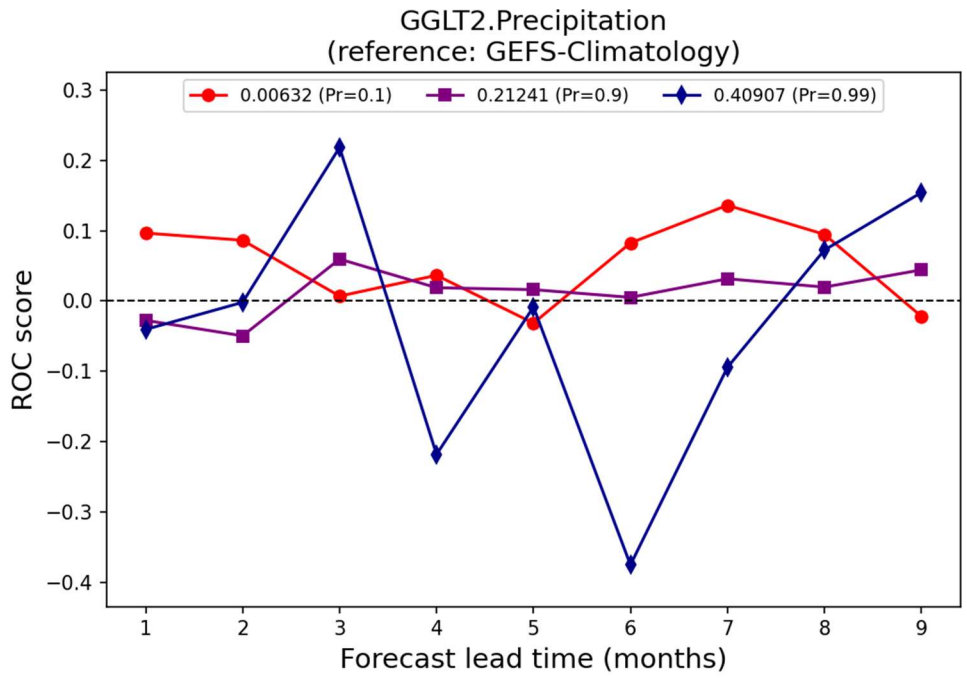

Figure 4-64: ROC scores of HEFS ensemble precipitation forecasts from GEFS-CFSv2against lead time at GAST2. The skill score is computed at 50, 90 and 99% quantile thresholds on postprocessed precipitation forecasts aggregated onto monthly intervals, with the climatological probabilities serving as the reference.

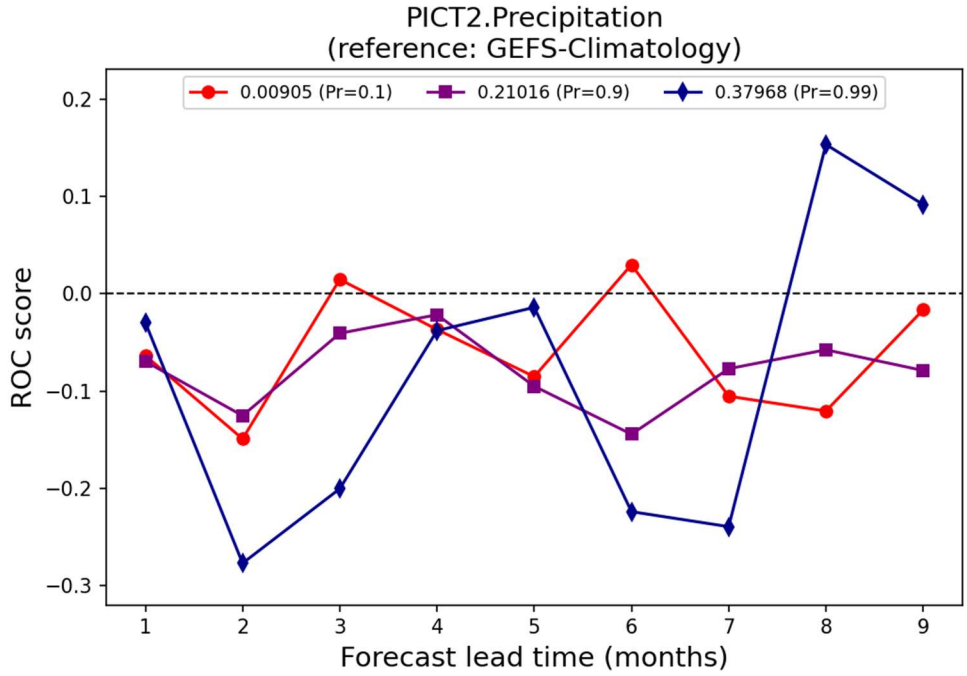

Figure 4-65: As Fig. 4-64, except at PICT2.

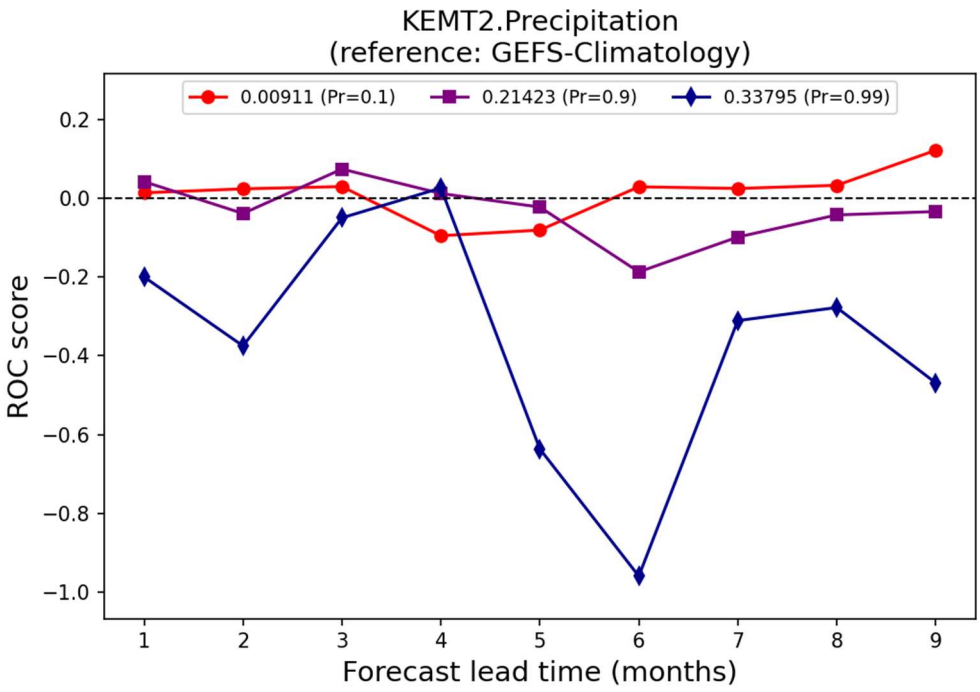

Figure 4-66: As Fig. 4-64, except at KEMT2.

The BSS and ROC scores for ensemble streamflow forecasts are shown in Figs. 4-56 - 4-61, and notable observations are summarized below.

- In terms of BSS, it appears that the streamflow forecasts from GEFS-CFSv2are broadly less skillful than those from GEFS-Climatology. An exception is PICT2, where for the bottom threshold (10%), BSS is positive at 1- and 2-month leads.
- ROC scores vary widely among thresholds and sites. At the southern-most site, KEMT2, ROC scores at the middle and bottom thresholds are consistently positive to month 9.
- At GAST2, ROC scores for the highest threshold are positive at 1- and 2-month leads, suggesting some discrimination skills at this range. At the middle threshold, the scores tend to increase with lead time, and stay above zero to 9-month lead, whereas those at the lowest threshold the scores are close to zero.
- At PICT2, ROC scores at the top and middle thresholds are positive at 1- and 2-month leads. At longer leads (6-9 months), the scores for the middle threshold are positive. By contrast, the scores are consistently positive at the lowest threshold.

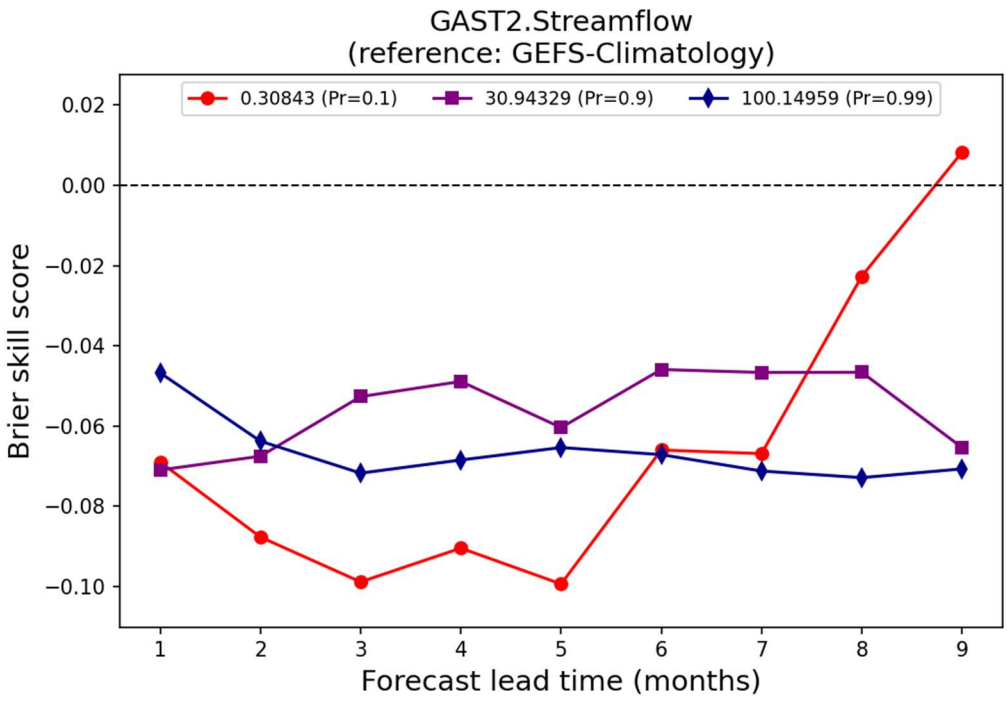

Figure 4-67: BSS of HEFS ensemble streamflow forecasts against lead time at GAST2. The skill score is computed at 10, 90 and 99% quantile thresholds on streamflow forecasts forced by GEFS-CFSv2 precipitation forecasts, with the ensemble streamflow forecasts forced by GEFS-Climatology serving as the reference. Note the streamflow forecasts are aggregated onto monthly intervals.

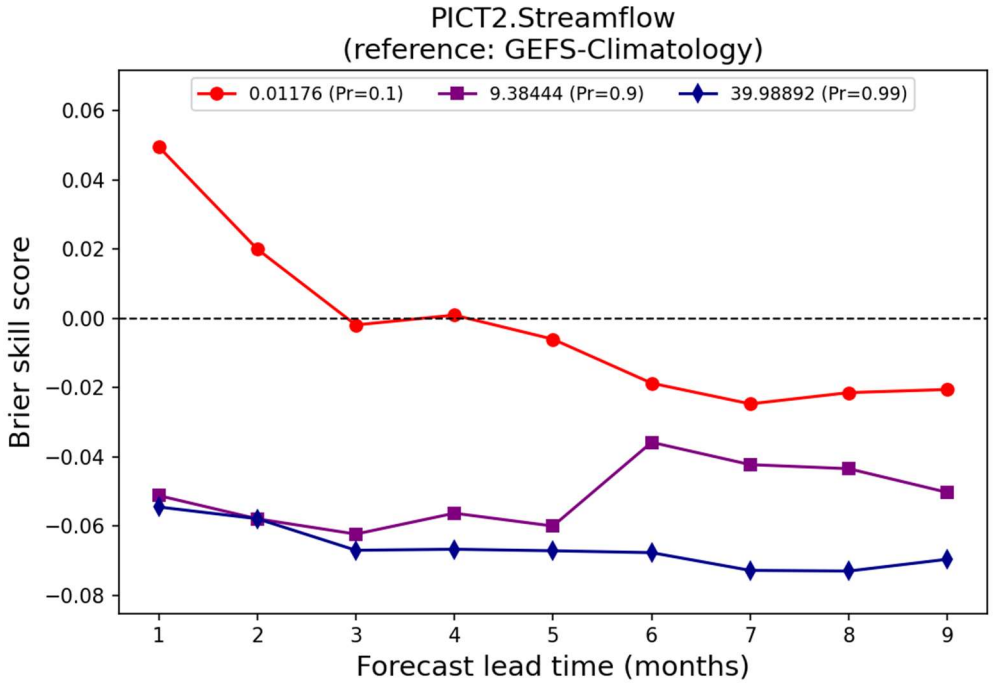

Figure 4-68: As Fig. 4-67, except at PICT2.

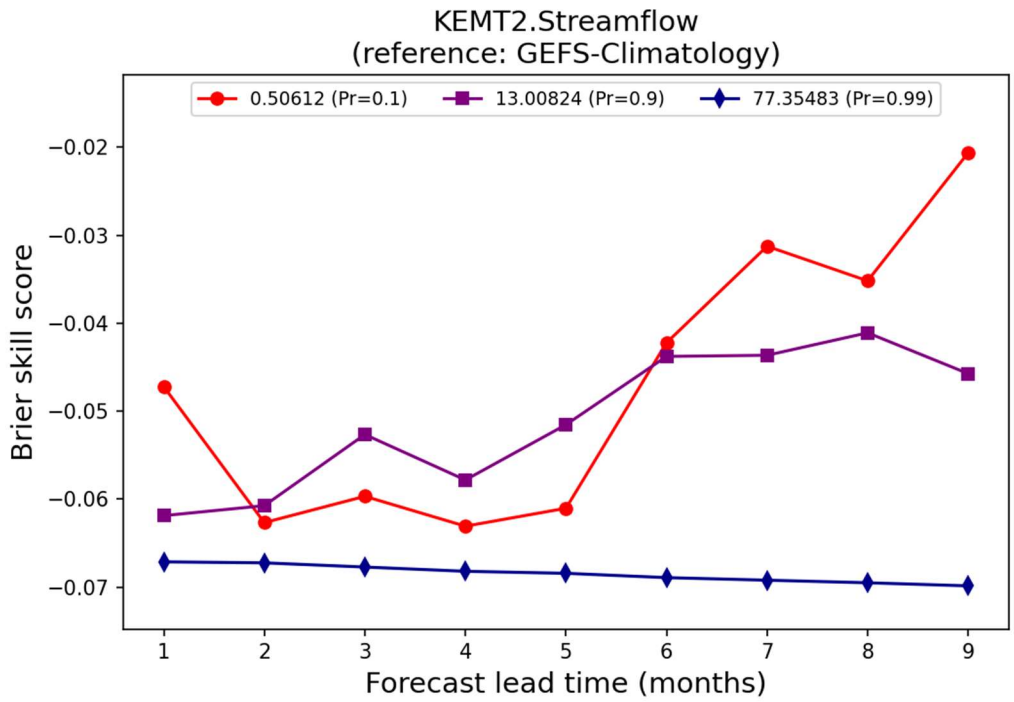

Figure 4-69: As Fig. 4-67, except at KEMT2.

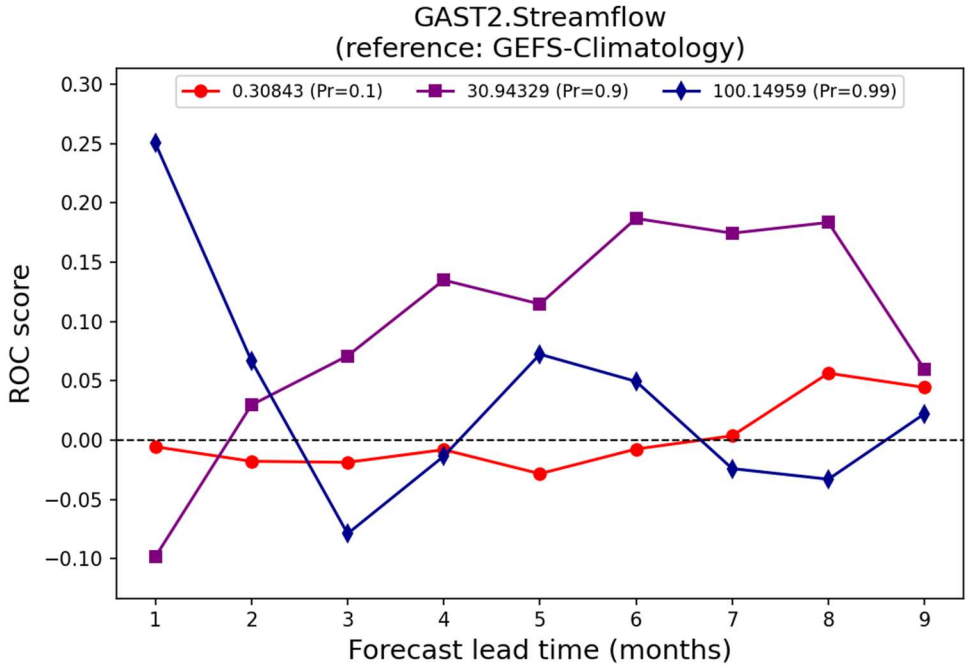

Figure 4-70: ROC scores of HEFS ensemble streamflow forecasts against lead time at GAST2. The scores are computed at 10, 90 and 99% quantile thresholds on streamflow forecasts forced by GEFS-CFSv2 precipitation forecasts, with the ensemble streamflow forecasts forced by GEFS-Climatology serving as the reference. Note the streamflow forecasts are aggregated onto monthly intervals.

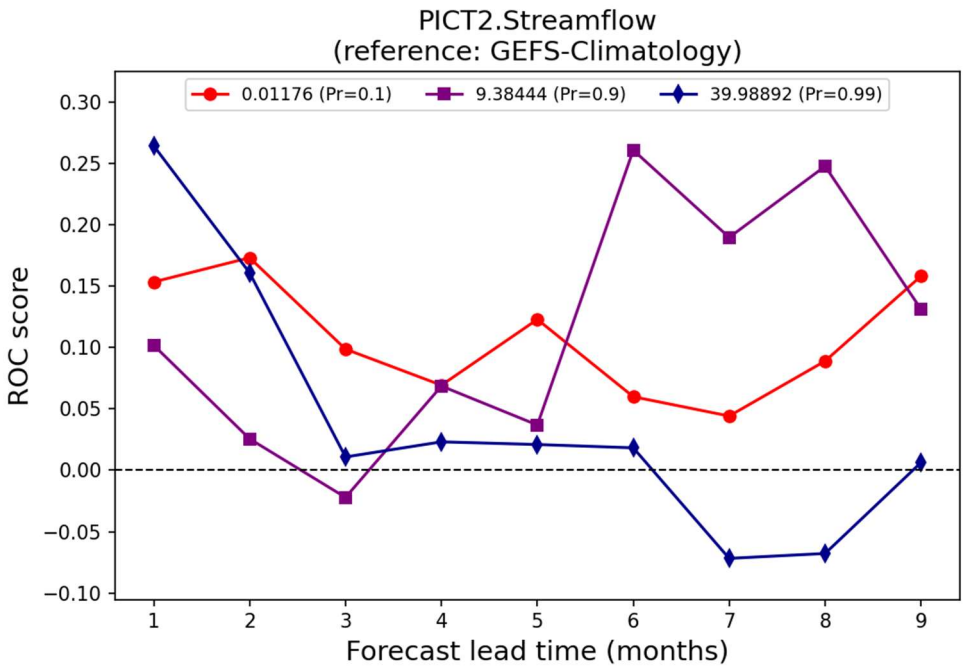

Figure 4-71: As Fig. 4-70, except at PICT2.

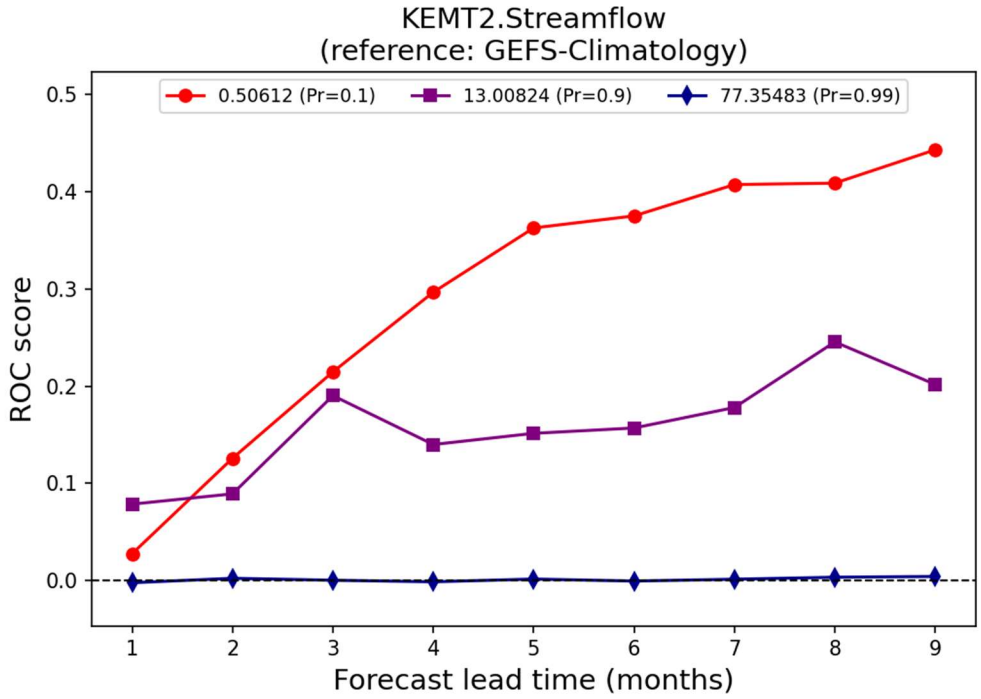

Figure 4-72: As Fig. 4-70, except at KEMT2.

4.4. HEFS ensemble streamflow forecasts to Lake Conroe:

As shown in Fig 4-62, Lake Conroe is situated along the West Fork of San Jacinto River to the north of Houston. It is a major water supply reservoir for the city of Houston. There are two NWS forecast points in the region: a) West Fork of San Jacinto River near Huntsville (SJHT2), and b) West Fork of San Jacinto River under Lake Conroe (LCTT2).

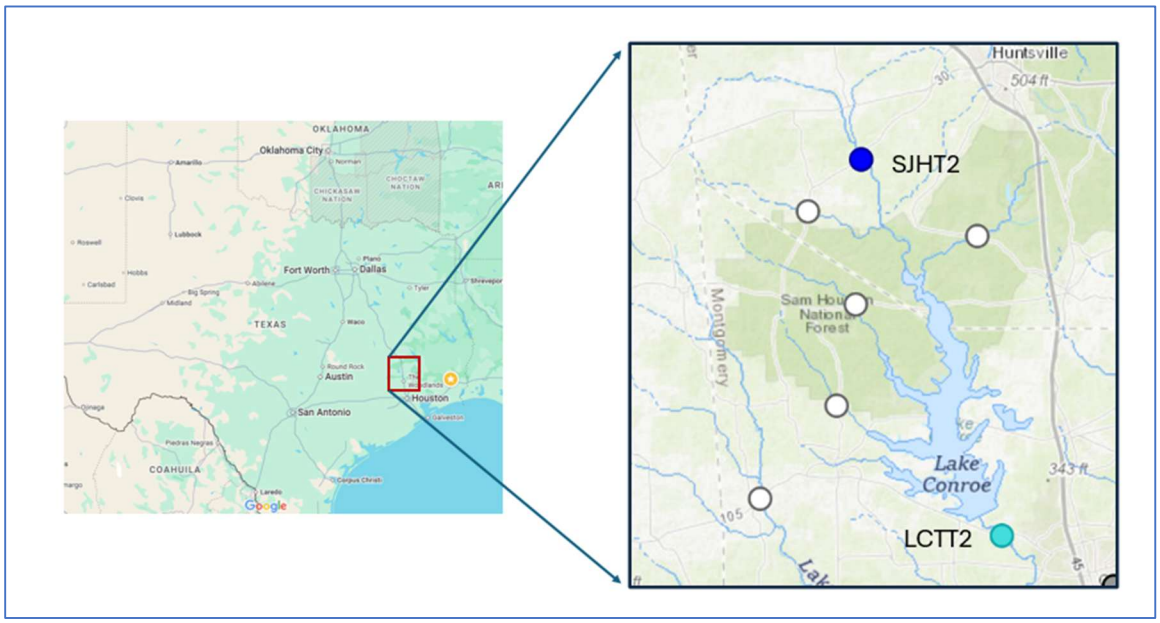

Figure 4-73: Map of West Fork of San Jacinto River and forecast/observation points near Lake Conroe. Two forecast points are located in the region, namely West Fork of San Jacinto River near Huntsville (SJHT2), and below Lake Conroe (LCTT2). SJHT is collocated with USGS station 08067548 and LCTT2 is collocated with USGS station 08067650.

To assist with the TWDB's pilot initiative at Lake Conroe, the UTA team collaborated with WGRFC to produce HEFS ensemble streamflow forecasts at the two forecast points over the following periods:

- 1 June – 30 September 2005
- 1 June – 30 September 2008
- 1 – 10 October 2011
- 1 – 5 October 2016
- 1 June – 30 September 2017
- 1 June – 30 September 2019

We examine the skills of HEFS ensemble streamflow forecasts for Hurricane Harvey in August 2017. Figures 4-63 and 4-64 compare the ensemble forecasts issued at 12z on 26 and 27 August for SJHT2 against observed daily mean discharge from USGS. It is evident that the forecasts issued two days ahead of the peak were unable to foresee the magnitude of the event, with 90% quantile of the peak less than a third the actual observed. The forecasts issued one day later are much higher, but the 90% quantile remains consistently below the observed.

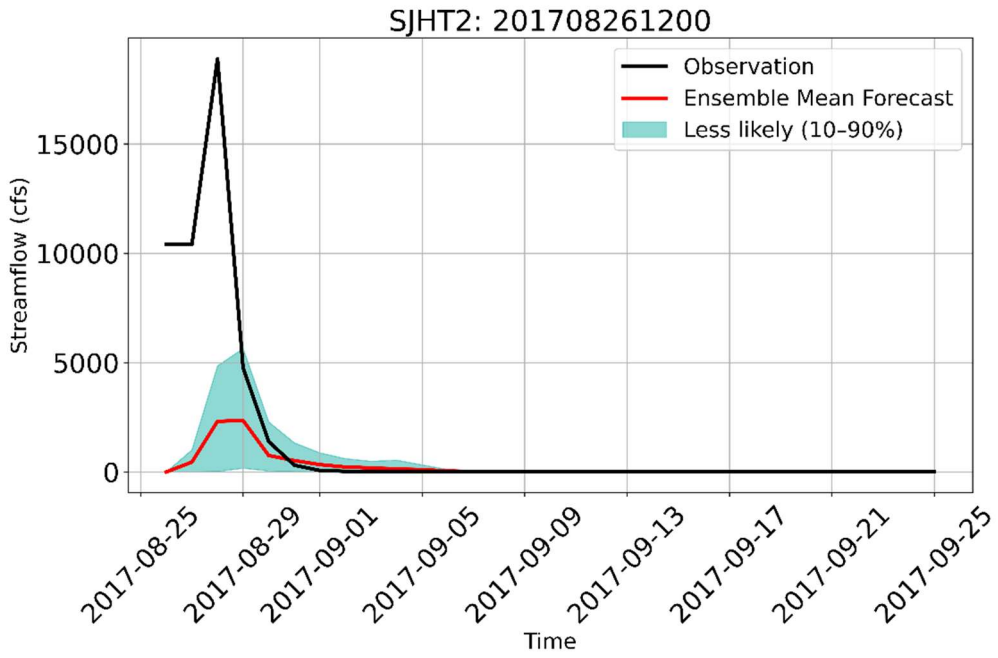

Figure 4-74: HEFS ensemble streamflow forecasts issued at 12z on 26 August 2017 and observed flow series from USGS.

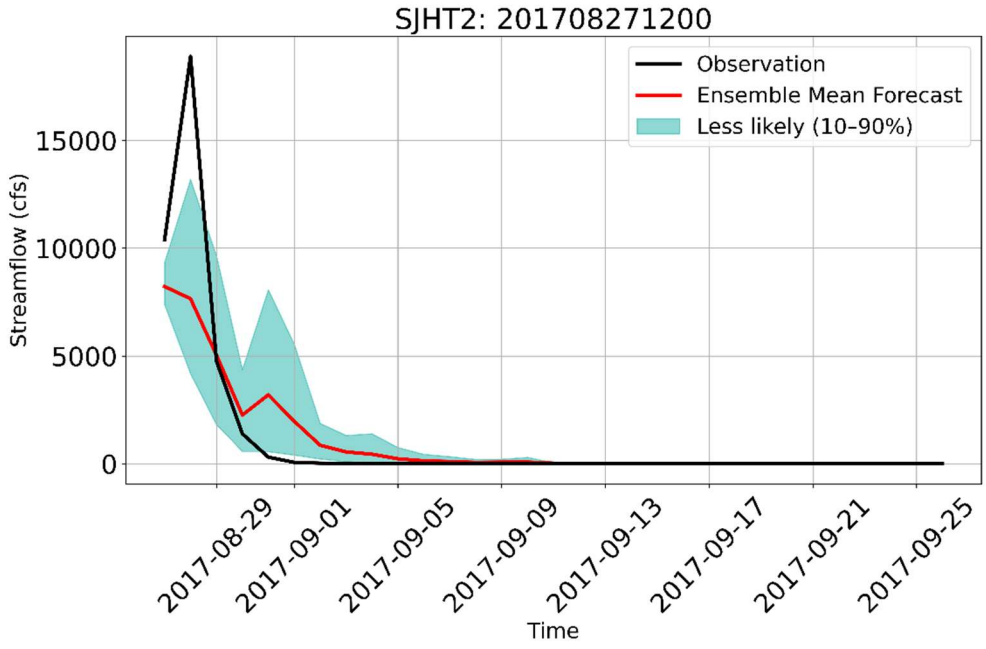

Figure 4-75: As Fig. 4-74, except for forecasts issued at 12z on 27 August 2017.

5. Summary and Recommendations

This project was launched as a part of the Texas FIRO initiative to determine skills in the HEFS ensemble forecasts at forecast points within the Texas FIRO Pilot, and more broadly to facilitate the application of the forecasts in reservoir operations in the state of Texas. Though NWS OWP performed baseline evaluation of HEFS forecasts, the earlier evaluation relied on an older version of the WGRFC model configuration, did not incorporate updated parameter values from the recent calibration done in 2022, and focused exclusively on skills at forecast points upstream of reservoirs for lead time of day 1-30. The current project addresses these limitations by expanding the scope of evaluation to determine skills of precipitation and streamflow forecasts at the extended range (> 30 days), and at reservoir inlets. In addition, the current project investigates the impacts from switching to the latest calibrated parameter set on streamflow simulations, and experimented with integrating alternative, S2S and seasonal forecasts as alternatives to resampled climatology at the extended range. Major findings from the project are summarized below.

Overall, Skills of HEFS Ensemble Precipitation and Streamflow Forecasts

Four HEFS configurations were created to determine the skills from different combinations of NWP forecast and climatology as forcing. These include Climatology, GEFS-Climatology, GEFS-S2S, and GEFS-CFSv2. Among these, Climatology mimics the legacy ESP by using resampled climatology alone as the forcing to drive the hydrologic model throughout the lead time range of HEFS (day 1-270) and serves as a key reference for gauging the skills in forecasts from alternative configurations. GEFS-Climatology is the operational default at WGRFC, whereas GEFS-S2S and GEFS-CFSv2 are experimental configurations.

Generally speaking, the analysis indicates that ensemble streamflow forecasts forced by the NWP precipitation forecasts are more skillful than those driven by resampled climatology in the medium range (days 1–12). When forecast variables are aggregated on monthly intervals, the skills extend to the S2S range (days 15–60). Among the three forecast configurations, namely, GEFS-Climatology, GEFS-S2S, and GEFS-CFSv2, the streamflow forecasts from the first two configurations are broadly comparable. There are signs that GEFS-S2S streamflow forecasts tend to be slightly more skillful at longer lead times, possibly pointing to the merit of employing GEFS S2S forecast, though the differences in forecast skills remain small. The analysis also suggests that forecasts from GEFS-CFSv2 are somewhat less skillful than GEFS-Climatology, consistent with observations by WGRFC that direct ingest of CFSv2 forecasts has either no, or at best marginal, impact on the skills of streamflow forecasts.

The lead time range at which the forecasts does vary among sites and metrics. In general, the HEFS streamflow forecasts tend to be more skillful at forecast points in the northern portion of the pilot domain, and those associated with larger drainage areas (e.g., GAST2, PICT2, BLNT2, STIT2), and less so at those in the southeast whose drainage areas are smaller (KEMT2, GGLT2, and GNGT2). At GAST2 and PICT2, for example, CRPSS of monthly flow forecasts are positive for the first two months, whereas at GGLT2 and GNGT2, CRPSS is barely positive/negative at 1-month lead. Further analysis of BSS and ROC scores yielded mixed results, with the former pointing to forecast skills at moderate/high flow thresholds (90 and 99% quantiles) in the first month, and the latter showing skills at longer lead times and across all thresholds for a majority of sites.

Note that though the use of NWP precipitation forecasts in the medium range help improve the skills of streamflow forecasts, the improvements tend to be limited to the first month. It is unsurprising that streamflow forecasts exhibit skills at lead times longer than those for precipitation, as watershed runoff response tends to dampen variability in precipitation and extend the time window of its impact. However, this does not explain the north-south gradient in the predictive skills. A detailed attribution analysis was not performed here, but it is likely that both hydroclimate and the size of drainage give rise to the gradient. Specifically, the contribution of less predictable, convective storms to the precipitation in the region may largely determine the overall predictive skills, and there is a possibility that these occurred more often in the southern portion of the pilot domain during the period of analysis (2000-2019).

Another notable observation is that the streamflow forecasts tend to be the most skillful for moderate events (>90% climatological quantile), and to a lesser for large events (>99% climatological quantile). Whereas for low flow (> 10% climatological quantile), the ensemble streamflow forecasts are broadly unskillful. This magnitude-dependent performance for streamflow forecasts is in direct contrast to that for precipitation forecasts, where skills for heavy/light precipitation tend to be lowest/highest. This inverse relationship is an indication of difficulties for the operational hydrologic model to reproduce both high and low flows.

Predictive Skills for Extreme Inflow Events

Two case studies were conducted illustrate the ability of HEFS to predict extreme inflow events that are particularly concerning to reservoir operators. The first is the June 2007 flood event that featured a sequence of convective storms that produced heavy rainfall in central and northern Texas, among which the storm in June 27 produced a heavy rainfall bullseye just to the southwest of the drainage to Lake Georgetown, resulting in near record water level in the reservoir by early July. Considering that heavy rainfall episodes tend to be rarer by late June over central Texas according to climatology (Nielson-Gammon et al, 2005), the occurrence of this event marks an aberration and challenges the potential of leveraging flood storage for summertime water supply in the region. The second event is Hurricane Harvey in late August 2017 that produced record inflow to Lake Conroe, a water supply reservoir to the north of Houston. While Lake Conroe does not provide a flood pool, accurate prediction of inflow and release for events such as Harvey would be key to emergency response for communities downstream of the reservoir.

Our analysis shows that the HEFS ensemble precipitation forecasts for both events were severely biased. For the June 2007 event, the 3-day forecasted precipitation amount upstream of Georgetown was less than one fifth of the observation, and the ensemble spread was too thin to contain the outcome. A primary contributor to the negative bias is the location error in the forecasted heavy rainfall center – the rainfall maximum in the GEFSv12 forecast was displaced to the north by 200 miles. This was further compounded by a negative bias in the forecasted magnitude of the peak rainfall rate. In addition, we suspect that the postprocessing mechanism in HEFS, the MEFP, played a role as it relies on a regression approach that tends to reduce the magnitude of ensemble mean (Kim and Seo, 2025).

It is also worth noting that, for this event, the ensemble forecasts of inflow to Georgetown exhibit even more severe negative bias, underscoring deficiencies in the operational hydrologic modeling system. Specifically, it appears that the SAC-SMA model that performs the water balance calculation was incapable of producing adequate runoff for this event. This raises the question on the adequacy of model calibration for high flow events.

For Hurricane Harvey, the 3-day forecast of inflow to Lake Conroe also featured severe negative bias and under-dispersed (with overly narrow ensemble spread). The 2-day forecast was much more skillful, though the magnitude remained low. Though precipitation forecasts were not evaluated, it is likely that much of the under-forecast in inflow was a result of under-forecast in precipitation.

Impacts of hydrologic model calibration

A comparison of streamflow simulations was performed using parameter values obtained from two earlier calibration projects completed in 2008 and 2022, with a focus on the performance of model simulations over major flood events. The results point to mixed impacts from the latest calibration effort in 2022. While some gains were seen in summary statistics such as correlation, the recent calibration appears to have introduced a consistent negative bias in the simulated peak discharge for major flood events. While for a few cases, using parameter values from the 2022 calibration resulted in overprediction of peak discharges, for a majority of events analyzed it led to overly depressed peaks and therefore degraded the accuracy of simulations.

These results suggest that the negative bias in SAC-SMA simulations is a systematic issue and not limited to the two events examined in the project, and that inadequate model calibration likely contributed to under-forecast of other major flood events. Though the calibration effort in 2022 was meant to improve the accuracy of SAC-SMA simulations, it nonetheless degraded the performance of the model for major events, possibly due to a lack of inclusion of metrics that measure the model errors for such events.

Recommendations:

On the basis of the findings, the UTA FIRO Pilot team makes the following recommendations.

Recommendation I: Improve forecast skills for anomalously large precipitation events in the medium range. This can be achieved through the following actions:

- Introduce alternative, advanced postprocessing algorithms to HEFS to alleviate the negative bias and under-spread in raw precipitation forecasts. UTA has developed an enhanced postprocessing scheme for the MEFP, namely the Conditional Bias Penalizing Regression (CBPR; Kim and Seo, 2025). CBPR uses an alternative method for estimating the parameters for the Mixed Meta-Gaussian Distribution (MMGD) that forms the basis for computing the predictive distribution of precipitation from pairs of observations and forecasts. In addition, the NOAA Physical Science Lab (PSL) proposed the Censored-Shifted Gamma Distribution (CSGD; Scheuerer and Hamill, 2015), which offers simpler, more robust mechanisms for establishing predictive distribution of precipitation. Recently, the UTA project team formulated an enhanced version of CSGD that uses Artificial Neural Network for training (ANN-CSGD; Ghazvinian et al., 2021, 2022) and demonstrated its prowess for producing reliable probabilistic forecasts of

heavy-to-extreme rainfall with limited training data. The UTA team is also working with PSL to develop a gridded version of MEFP that implements the CSGD, a gridded version of the MMGD (Zhang et al., 2017), and an enhanced version of Schaake Shuffle (Wu et al., 2017). Integrating the output from this gridded MEFP will help determine its efficacy in remedying spatial displacement-related forecast biases.

- Explore alternative precipitation forecasts as input to HEFS. The European Center for Medium-range Weather Forecasts (ECMWF) now produces high-resolution ensemble forecasts (9-km) out to day 15, which have been shown to perform favorably against the GFS for many variables. The NWS has been producing high-resolution forecasts for the US with the High-Resolution Rapid Refresh (HRRR) system, HRRR produces 3-km forecasts out to 48 h using a convection permitting model and proves more capable in resolving precipitation induced by convection. Note that the ECMWF maintains a reforecast archive that can be leveraged for training the postprocessing systems. HRRR does not have a reforecast archive, but a real-time archive is available from 2014.

<u>Recommendation II: Improve the calibration of NWS hydrologic and routing models to better capture the magnitude of inflow during flood events.</u>

- Recalibrate SAC-SMA for forecast points in the pilot domain to address the negative bias evident in the SAC-SMA simulations for major flooding events. Metrics such as mean volumetric biases for annual flood episodes may be used in addition to those employed in earlier calibration efforts, such as bias, correlation, NSE and KGE for long-term simulations. If warranted, the routing model may also undergo calibration.
- Quality control reservoir inflow estimates. The reservoir inflow estimates were produced through mass balance calculations that can be error-prone due to a variety of factors such as water surface gradient with reservoir during passage of flood waves, inaccuracy in estimates of release, withdrawal and evaporation, and wind-induced waves. A more rigorous comparison between rainfall time series and the inflow estimates will help isolate periods when the estimates would be of sufficient quality to serve as reference for model calibration and forecast verification.
- Determine optimal precipitation products for calibration and verification. In this project we use the NWS AORC precipitation product as the ground truth for verification, though NWS used the WGRFC Mean Areal Precipitation (MAP) product in estimating the MEFP parameters. The project team has found that the AORC product in the years prior to the release of the Stage-IV product has suffered bias and large errors due to issues in the North American Land Data Assimilation System – II (NLDAS-II; Xia et al., 2012) that serves as the source. The team has since developed a bias-corrected AORC product following the approach of Zhang et al., (2010) that underwent limited evaluation and was shown to outperform the raw AORC dataset (Zhang et al., 2025). It is recommended that the WGRFC MAP, AORC and the bias-corrected AORC undergo additional assessment to determine the optimal product for SAC-SMA calibration, MEFP parameter estimation, and forecast verification.

<u>Recommendation III**:** Investigate alternative forecast products as forcing to HEFS beyond the medium range.</u>

As noted earlier, neither GEFS S2S forecasts (out to day 35) nor the CFSv2 forecasts produced appreciable gains in forecast skills when ingested into HEFS through the MEFP. However, it

remains possible that signals in these forecasts have not been fully exploited due to a lack of spatial preciseness in the extended range forecasts by NWP and General Circulation Models. Approaches such as forecast analogs, use of large, multi-model ensemble, and weather generators may help further improve forecast skills for impactful events at the S2S range. Several forecast improvement efforts are already underway at FIRO partnering organizations to explore these techniques and the outcomes, including a UTA initiative aimed to develop analog ensemble forecasts for the S2S range and an effort by the PSL to examine forecast-guided resampling of ensemble traces for weeks 3–5.

Recommendation IV**:** Partner with reservoir operators to add and refine metrics for forecast evaluation and improvements to facilitate consistent use of forecasts in operation decisions. One of the key objectives of FIRO Pilot is to collaborate with reservoir operators to improve forecast skills and delivery, with the ultimate aim of making the forecasts a regular ingredient in reservoir operations. The metrics employed in foregoing evaluations have seen frequent use by forecasters. Yet, reservoir operators' exposure to them has been limited to date, and their specific ability to reflect risks associated with decisions has yet to be demonstrated. It will be helpful to work closely with reservoir operators to identify scenarios where the impacts from potential failures of forecasts can be gauged and accounted for.

For example, the 99% climatological quantiles of daily, weekly, and monthly precipitation or flow were used as the thresholds of rare events in computing BSS and ROC scores. Both quantile threshold and the accumulation window may be fine-tuned to reflect the risk and risk tolerance of reservoir operators. The National FIRO Program now uses Critical Success Index and Dry Forecast Failure ratio as metrics in its screening of reservoirs. The metrics, though intended for deterministic forecasts, help prepare reservoir operators prepare for potential high flow events that pose risks to dam safety and downstream communities, and their inclusion in future analysis will be recommended.

Acknowledgement

Many individuals contributed to the launch of FIRO Pilot, and to the completion of the hindcast and validation. These include Kris Lander, Andrew Philpot and Frank Bell at WGRFC, Jerry Cotter, Max Strickler, John Hunter and other colleagues from the USACE Fort Worth District, Aaron Abel, August Dreyer, Peyton Lisenby and Chris Higgins at Brazos River Authority, and John Zhu and Nelun Fernando at the Texas Water Development Board. Professor Dong-Jun Seo at UTA and James Brown at NWS provided guidance to the project team on forecast evaluation. Zhengtao Cui set up the CHPS system on UTA servers and assisted with troubleshooting and hindcast experiments.

References

Anderson, E.A., 1973. *National Weather Service river forecast system: Snow accumulation and ablation model* (Vol. 17). US Department of Commerce, National Oceanic and Atmospheric Administration, National Weather Service.

Burnash, R.J., 1973. A generalized streamflow simulation system: Conceptual modeling for digital computers. US Department of Commerce, National Weather Service, and State of California, Department of Water Resources.

Curtis, D.C. and Schaake, J.C., 1979. The NWS extended streamflow prediction technique. ASCE.

Day, G.N., 1985. Extended streamflow forecasting using NWSRFS. Journal of Water Resources Planning and Management, 111(2), pp.157-170.

Demargne, J., Wu, L., Regonda, S.K., Brown, J.D., Lee, H., He, M., Seo, D.J., Hartman, R., Herr, H.D., Fresch, M. and Schaake, J., 2014. The science of NOAA's operational hydrologic ensemble forecast service. BAMS, 95(1), pp.79-98.

Ghazvinian, M., Zhang, Y., Seo, D.J., He, M. and Fernando, N., 2021. A Novel Hybrid Artificial Neural Network-Parametric Scheme for Postprocessing Medium-Range Precipitation Forecasts. Advances in Water Resources, p.103907.

Ghazvinian, M., Zhang, Y., Hamill, T.M., Seo, D.J. and Fernando, N., 2022. Improving probabilistic quantitative precipitation forecasts using short training data through artificial neural networks. *Journal of Hydrometeorology*, *23*(9), pp.1365-1382.

Kim, S., H. Sadeghi, R. A. Limon, D.-J. Seo, A. Philpott, F. Bell, J. Brown, K. He, 2018. Ensemble streamflow forecasting using short- and medium-range precipitation forecasts for the Upper Trinity River Basin in North Texas via the Hydrologic Ensemble Forecast Service (HEFS). J. of Hydromet. 19(9), pp.1467-1483.

Kim, S. and Seo, D.J., 2025. Improving Ensemble Precipitation and Streamflow Forecasts for Large Events with the Conditional Bias-Penalized Regression-Aided Meteorological Ensemble Forecast Processor. Weather and Forecasting, 40(6), pp.959-975.

Nielsen-Gammon, J.W., Zhang, F., Odins, A.M. and Myoung, B., 2005. Extreme rainfall in Texas: Patterns and predictability. Physical Geography, 26(5), pp.340-364.

Regonda, S., and Seo, D.-J. 2008. Statistical post processing streamflow ensembles to improve reliability over a wide range of time scales. 2nd CPPA PIs Meeting, Silver Spring, MD, NOAA, http://vintage.joss.ucar.edu/joss_psg/meetings/Meetings_2008/CPPA/poster/RegondaPoster.pdf.

Saha, S., Moorthi, S., Wu, X., Wang, J., Nadiga, S., Tripp, P., Behringer, D., Hou, Y.T., Chuang, H.Y., Iredell, M. and Ek, M., 2014. The NCEP climate forecast system version 2. *Journal of climate*, *27*(6), pp.2185-2208.
Scheuerer, M. and Hamill, T.M., 2015. Statistical postprocessing of ensemble precipitation forecasts by fitting censored, shifted gamma distributions. *Monthly Weather Review*, *143*(11), pp.4578-4596.

TWDB 2020. Forecast-informed Reservoir Operations (FIRO) and Water Resources Management in Texas and Oklahoma. Available at https://www.twdb.texas.gov/publications/reports/other_reports/doc/TWDB_UTA_NIDIS_forecasts_workshop_report.pdf

Wu, L., Seo, D.J., Demargne, J., Brown, J.D., Cong, S. and Schaake, J., 2011. Generation of ensemble precipitation forecast from single-valued quantitative precipitation forecast for hydrologic ensemble prediction. *Journal of hydrology*, *399*(3-4), pp.281-298.

Wu, L., Zhang, Y., Adams, T., Lee, H., Liu, Y. and Schaake, J., 2018. Comparative evaluation of three Schaake shuffle schemes in postprocessing GEFS precipitation ensemble forecasts. *Journal of Hydrometeorology*, *19*(3), pp.575-598.

Xia, Y., Mitchell, K., Ek, M., Sheffield, J., Cosgrove, B-, Wood, E., Luo, L., Alonge, C., Wei, H., Meng, J. and Livneh, B-, 2012. Continental-scale water and energy flux analysis and validation for the North American Land Data Assimilation System project phase two (NLDAS-2): 1. Intercomparison and application of model products. Journal of Geophysical Research: Atmospheres, 117(D3).

Zhang, Y., Reed, S. and Kitzmiller, D., 2011. Effects of retrospective gauge-based readjustment of multisensor precipitation estimates on hydrologic simulations. Journal of Hydrometeorology, 12(3), pp.429-443.

Zhang, Y., Wu, L., Scheuerer, M., Schaake, J. and Kongoli, C., 2017. Comparison of probabilistic quantitative precipitation forecasts from two postprocessing mechanisms. *Journal of Hydrometeorology*, *18*(11), pp.2873-2891.

Zhang, Y., Frakheddine, S. and Hayes, J., 2025. A model-based Investigation of Streamflow Trends in Upper Brazos River Basin. Contracted Report to the Texas Water Development Board.