



FINAL REPORT

Hydrologic Risk Forecaster (Hydro RF): Demonstration and Validation of Forecasting Model for Flooding and Infrastructure Risk on Military Lands

Resource Conservation and Resiliency Projects

RC19-5082

Daniel Gambill | Engineering Research and Development Center-Construction
Engineering Research Laboratory (ERDC-CERL)

August 2022

DISTRIBUTION STATEMENT A. Approved for public release: distribution unlimited.

REPORT DOCUMENTATION PAGE					Form Approved OMB No. 0704-0188	
The public reporting burden for this collection of information is estimated to average 1 hour per response, including the time for reviewing instructions, searching existing data sources, gathering and maintaining the data needed, and completing and reviewing the collection of information. Send comments regarding this burden estimate or any other aspect of this collection of information, including suggestions for reducing the burden, to Department of Defense, Washington Headquarters Services, Directorate for Information Operations and Reports (0704-0188), 1215 Jefferson Davis Highway, Suite 1204, Arlington, VA 22202-4302. Respondents should be aware that notwithstanding any other provision of law, no person shall be subject to any penalty for failing to comply with a collection of information if it does not display a currently valid OMB control number. PLEASE DO NOT RETURN YOUR FORM TO THE ABOVE ADDRESS.						
1. REPORT DATE (DD-MM-YYYY) 31-08-2022		2. REPORT TYPE ESTCP Final Report			3. DATES COVERED (From - To) March 2019 - Sept 2023	
4. TITLE AND SUBTITLE Hydrologic Risk Forecaster (Hydro RF): Demonstration and Validation of Forecasting Model for Flooding and Infrastructure Risk on Military Lands					5a. CONTRACT NUMBER	
					5b. GRANT NUMBER	
					5c. PROGRAM ELEMENT NUMBER	
6. AUTHOR(S) Daniel R. Gambill, Sean Matus, Sandra LeGrand, John Eylander, Adam Cisco, Brendon Hoch, Alex Gallagher, Francina Dominguez					5d. PROJECT NUMBER RC19-5082	
					5e. TASK NUMBER	
					5f. WORK UNIT NUMBER	
7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES) USACE ERDC Construction Engineering Research Laboratory 2902 Newmark Dr, Champaign, IL 61822 USACE ERDC Cold Regions Research and Engineering Laboratory 72 Lyme Rd, Hanover, NH 03755					8. PERFORMING ORGANIZATION REPORT NUMBER RC19-5082	
9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES) Environmental Security Technology Certification Program (ESTCP) Office of the Deputy Assistant Secretary of Defense (Energy Resilience & Optimization) 3500 Defense Pentagon, RM 5C646 Washington, DC 20301-3500					10. SPONSOR/MONITOR'S ACRONYM(S) ESTCP	
					11. SPONSOR/MONITOR'S REPORT NUMBER(S) RC19-5082	
12. DISTRIBUTION/AVAILABILITY STATEMENT DISTRIBUTION STATEMENT A. Approved for public release: distribution unlimited.						
13. SUPPLEMENTARY NOTES						
14. ABSTRACT The objective of this work is to create a forecasting model that can provide DoD with near-term flood hazard assessments for soldier safety, critical infrastructure failure, and installation readiness (e.g. restrictions in access to training areas, reliability of critical resources such as potable water intake and pumping stations, training range management and degradation). Improved decision-making capability will allow for improved soldier safety, lower costs due to infrastructure failure or in-situ warning systems, increased installation resiliency by mitigating impacts to critical infrastructure, and meeting mission parameters by limiting unscheduled training downtime.						
15. SUBJECT TERMS Infrastructure, climate impacts on Installations, Installation Resilience, precipitation, flooding, Hydrologic Risk Forecaster, Hydro RF, Forecasting Model, Infrastructure Risk						
16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT UNCLASS	18. NUMBER OF PAGES 117	19a. NAME OF RESPONSIBLE PERSON Daniel R Gambill	
a. REPORT UNCLASS	b. ABSTRACT UNCLASS	c. THIS PAGE UNCLASS			19b. TELEPHONE NUMBER (Include area code) 217-373-3373	

TABLE OF CONTENTS

1.0	INTRODUCTION	1
1.1	BACKGROUND.....	1
1.2	OBJECTIVE OF THE DEMONSTRATION	3
1.3	REGULATORY DRIVERS.....	3
2.0	TECHNOLOGY DESCRIPTION	4
2.1	TECHNOLOGY OVERVIEW	4
2.1.1	Atmospheric Component	5
2.1.2	Hydrology and Hydraulic Component.....	9
2.1.3	Risk Communication Component.....	10
2.2	ADVANTAGES AND LIMITATIONS OF THE TECHNOLOGY	11
3.0	PERFORMANCE OBJECTIVES	13
3.1	QUANTITATIVE PERFORMANCE OBJECTIVES	15
3.1.1	Accurate probability precipitation forecasts	15
3.1.2	Accurate probability hydrologic forecasts	16
3.1.3	Forecasted risk agreement with historic conditions.....	19
3.1.4	Downscaling of ensemble members with good computational efficiency	20
3.2	QUALITATIVE PERFORMANCE OBJECTIVES.....	20
3.2.1	Risk Report ease of use.....	20
4.0	SITE DESCRIPTION	22
4.1	FORT HOOD	22
4.1.1	Location and History.....	22
4.1.2	Site Characteristics.....	23
4.2	FORT LEONARD WOOD	26
4.2.1	Location and History.....	26
4.2.2	Site Characteristics.....	29
4.3	FORT CARSON	32
4.3.1	Location and History.....	32
4.3.2	Site Characteristics.....	34
5.0	TEST DESIGN	38
5.1	CONCEPTUAL TEST DESIGN	38
5.2	BASELINE CHARACTERIZATION AND PREPARATION.....	39
5.3	DESIGN AND LAYOUT OF TECHNOLOGY AND METHODOLOGY COMPONENTS	39
5.3.1	Water Depth Sensors.....	39
5.3.2	Atmospheric Component	39
5.3.3	Hydrology Models	43
5.3.4	Hydraulics Models.....	61
5.3.5	Downscaling Stress Test	62
5.3.6	Risk Report communication strategie.....	68
6.0	PERFORMANCE ASSESSMENT	71
6.1	Accurate probability forecasts for precipitation.....	71
6.2	Accurate probability forecasts for streamflow time series.....	78
6.3	Forecasted risk agreement with historic conditions	94
6.4	Downscaling of ensemble members with good computational efficiency.....	90
6.5	Risk report ease of use	94

7.0	COST ASSESSMENT	107
8.0	IMPLEMENTATION ISSUES	111
9.0	FUTURE WORK.....	112

1.0 INTRODUCTION

Military installations consist of built infrastructure and range lands that must interact with natural riverine systems and mitigate the impacts of stormwater. In May 2016 the primary water pumping station intake at Fort Leonard Wood, MO was flooded by high water due to extreme rainfall (Sanders, 2017). By the time the pumping station was back in service, the installation had only 36 hours of potable water available. Extreme rainfall and flood events can also pose hazardous to military personnel. Roads across Camps Casey, Hovey, and Mobile in South Korea were made impassable in July 2011 due to heavy rains and resulting mudslides (Rabiroff, 2011). Road conditions in the area were downgraded to red, meaning only military vehicles on emergency or essential business could be driven, while the 8th Army HQ in Yongsan Garrison near Seoul was shut down for a time. In June of 2016, nine US Army soldiers drowned at Fort Hood, TX, when their LMTV was swept downstream while attempting to cross an at-grade LWC during heavy flooding (Stanglin, 2016). On April 3, 2017, another soldier was killed at Fort Hood when his car was swept away during a flash flood, and a civilian diver drowned a few days later while looking for the original victim (Myers, 2017).

The objective of this work is to create a forecasting model that can provide DoD with near-term flood hazard assessments for soldier safety, critical infrastructure failure, and installation readiness (e.g. restrictions in access to training areas, reliability of critical resources such as potable water intake and pumping stations, training range management and degradation). Improved decision-making capability will allow for improved soldier safety, lower costs due to infrastructure failure or in-situ warning systems, increased installation resiliency by mitigating impacts to critical infrastructure, and meeting mission parameters by limiting unscheduled training downtime.

1.1 BACKGROUND

Transportation infrastructure (roads, bridges, and LWCs), potable water systems, sanitary systems, treatment plants, and training ranges are often built in or near flood prone areas and are thus susceptible to flood events. DoD (2018) found 706 DoD real property sites are susceptible to potential flooding risk while 763 are susceptible to wind damage caused by storm events. The top five asset categories most reported as having one or more effects in the past were: airfield operations, followed by transportation infrastructure, energy infrastructure, training/range facilities, and water/wastewater systems. Installations such as Fort Riley, KS, are susceptible to severe storm events and resulting flash floods (Fiedler, 2013). Such extreme events can not only damage infrastructure near riverine systems, but can also destroy roads, sewers, and buildings in urbanized environments. These events can be so dangerous that they become deadly (Paige, 2018). Motor vehicle related deaths in floods account for more than half of all flood fatalities in the United States. Males, in particular, display a dangerous tendency to try to drive through floodwaters (Drobot et al., 2007)

The NWS currently operates an end-to-end HEFS (Demargne et al., 2014) as part of the AHPS (McEnery et al. 2005). AHPS produces ensemble streamflow predictions at over 4000 USGS streamflow gauges across the United States with plans to further expand. Current warning systems, such as these, are limited to only so many locations that have the necessary USGS streamflow

gauge. The NWS, with assistance from its RFCs, provides more encompassing flood warnings and watches that cover greater spatial and temporal scales. These watches and warnings can be issued for an entire county and for durations ranging from hours to days. The lack of applicability to remote locations and specificity with regards to timing and location makes the flood warning impersonal. If people do not think that emergency information was meant for them (the "it-can't-happen-to-me" syndrome), they likely will ignore it (Mileti, 1995). Also, repeated occurrences of false alarms cause the public to become desensitized to flood warnings over time (Downing, 1977). Other warning systems, such as water depth gauges, are reactive to present conditions. Reactive safety measures have lags between threat recognition and response, thus putting soldiers in danger (Parker et al., 2009). In-situ warning systems also rely on individual soldiers and on-site leadership making real-time decisions, sometimes without proper training (DoA, 2016).

There is a need for proactive safety measures that can anticipate future conditions and provide decision makers with the information needed to implement a precautionary response before a crisis, such as the Fort Hood deaths, develops. Near-term (6–72 hours) precipitation and streamflow forecasting is needed to assess risk, including soldier safety, predicted restrictions in access to training areas, reliability of critical resources such as potable water intake and pumping stations, and failure of critical infrastructure due to flooding. Near-term risk assessments (e.g. go/no go for remote LWC structures, flood inundation of potable water pumping stations, sanitary treatment plants and other critical infrastructure, and the ability to optimize training land usage with regards to accessibility and usability) would give military installations the capability to minimize risk associated with hydrologic events through predictions that are both more accurate and with longer lead times than current practice. The goal of the proposed work is to demonstrate and validate a tool for Range Control, DPW, DES, and/or other decision makers that uses near-term forecasts to limit flooding hazards that claim soldiers' lives and cause infrastructure failures. The safety risk to soldiers can be limited by optimizing the location and timing of training, while the risk of infrastructure failure can be limited with preemptive mitigation.

Hydro RF is a simulation and prediction tool that uses regional-scale, near-term deterministic and ensemble precipitation forecasts to predict subsequent hydrologic responses for riverine and stormwater infrastructure systems. The Hydro RF atmospheric module compiles near-term regional rainfall predictions distributed by operational weather data centers. The operational atmospheric data spans synoptic and mesoscale spatial scales. The rainfall data can be used to forecast individual events in the near-term. Hydro RF ingests high-resolution deterministic precipitation forecasts, as well as coarser probabilistic ensemble forecasts that are downscaled to the same high-resolution as input for Hydro RF's land surface module. The land surface module is a series of watershed-specific hydrologic and hydraulic models, which forecast individual streamflow/flooding events and the hydraulic responses at critical infrastructure. Hydro RF uses flow rate time series to generate risk reports at various outlooks (e.g. 12-hour lead time). Within each report, Hydro RF isolates if an event is in the forecast and the variance of that event across model members is conveyed in magnitude and arrival time through probability density functions (PDFs). The magnitude of the event is communicated as a probability of being within prescribed thresholds unique to a site (i.e. water depth and velocity at a LWC).

1.2 OBJECTIVE OF THE DEMONSTRATION

The purpose of this project is to demonstrate and validate the ERDC-CERL Hydro RF. We will achieve this through:

- 1) Demonstration of the Hydro RF model at three DoD installations representing three climatic regions (e.g. southeast, southwest, and midwest).
- 2) Validation of downscaling techniques and stress test of probabilistic ensemble forecasts.
- 3) Method validation for a range of input data quality and several applications/infrastructure types, such as LWCs, potable water intakes, sanitary water treatment infrastructure, and/or range training infrastructure.
- 4) Validation of the usability/actionability of flood risk reports based on feedback from installation decision makers.

1.3 REGULATORY DRIVERS

Heat waves, droughts, wildfires, and floods are straining already fragile ecosystems, hindering the ability to respond and recover from them, and affecting the Army missions on installations. Ensuring resilient installations, through energy and water security, improved efficiencies and diligent management of resources, reduces the vulnerabilities. To reduce vulnerabilities and sustain installation missions, Army Directive 2017-07 sets a requirement to secure critical missions by providing necessary energy and water for a minimum of 14 days (Speer, 2017). Key requirements include assured access to resource supplies, reliable infrastructure conditions, and effective system operations. Unfortunately, extreme storm and flood events continue to disrupt and threaten installation missions. Army Regulation 525–26 requires that the Army provide physical protection of critical infrastructures and key assets, including Army installations and mission-supporting holdings. The Army is mandated with the tasks of analyzing infrastructure risk, determining trends, and developing both short- and long-term infrastructure risk mitigation and management strategies.

2.0 TECHNOLOGY DESCRIPTION

2.1 TECHNOLOGY OVERVIEW

Hydro RF is a simulation and prediction tool which uses regional-scale, near-term deterministic and ensemble precipitation forecasts to predict subsequent hydrologic responses for riverine and stormwater infrastructure systems. Hydro RF was developed in 2016-2018 at Fort Hood, Texas. The main models comprising the Hydro RF framework include UCAR's WRF (Skamarock et al. 2008), HEC-HMS (USACE, 2000), HEC-RAS (USACE, 2010) and a probabilistic risk assessment methodology (Figure 2.1).

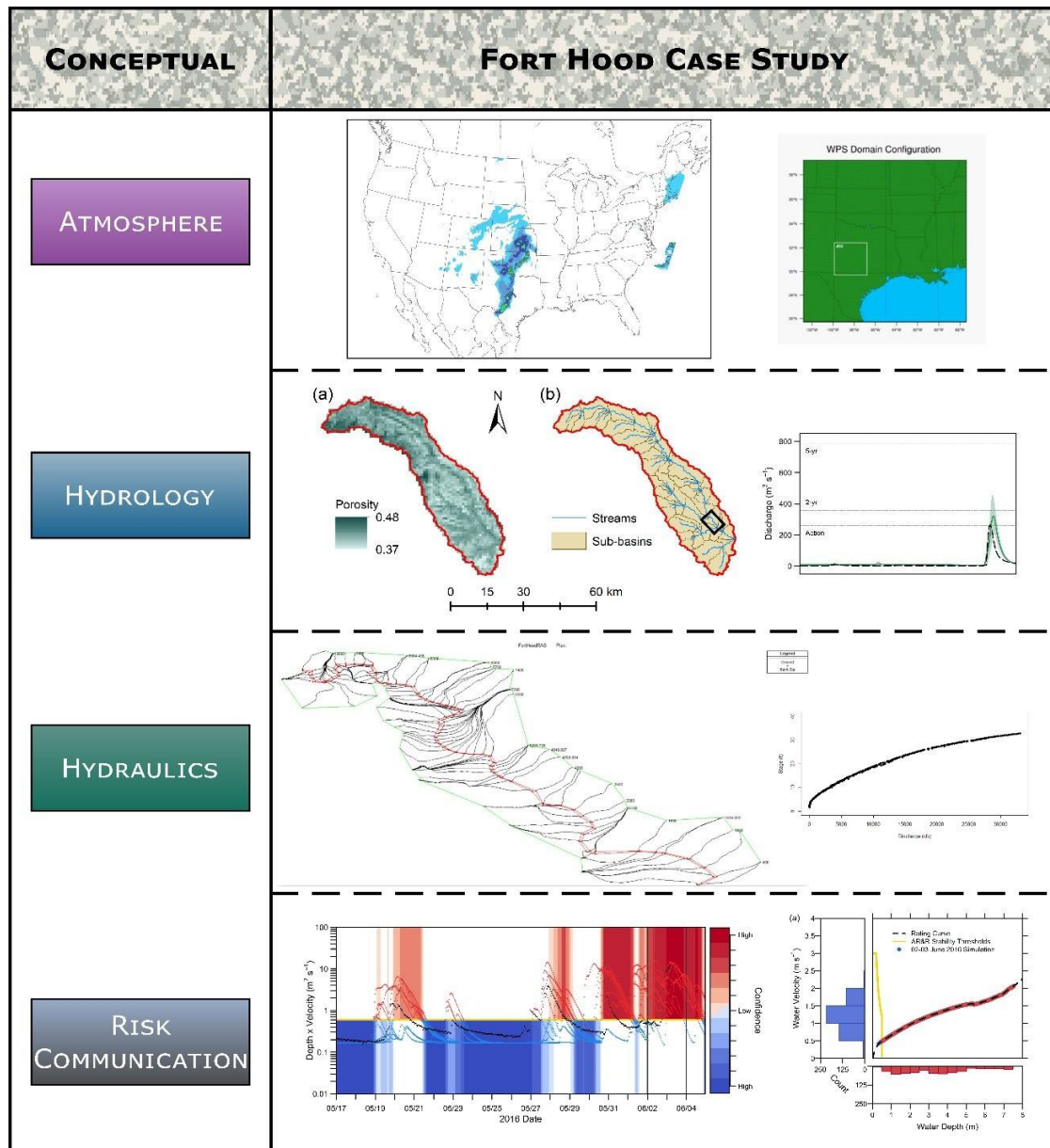


Figure 2.1: Hydro RF components

Hydro RF has the capability to run input data from deterministic forecasts or probabilistic forecast ensembles. Figure 2.2 illustrates data flows between the various components of the Hydro RF system and the various weather forecasts included. Inputting multiple deterministic and ensemble weather model output from different agencies will enable Hydro RF to incorporate uncertainty in initial conditions and model physics. Each weather forecast is represented as a unique member (or several unique members for the case of input ensembles) in a conglomerated Hydro RF ensemble. Each Hydro RF weather ensemble member is utilized as forcing for a single hydrology and hydraulic model simulation. The Hydro RF workflow produces an ensemble of streamflow predictions with an equal number of members to the conglomerated ensemble. The variability in atmospheric forcing across the Hydro RF ensemble trickles down through the hydrologic and hydraulic models in parallel to produce a swath of forecasted streamflow that objectively quantifies the probability of occurrence of individual events.

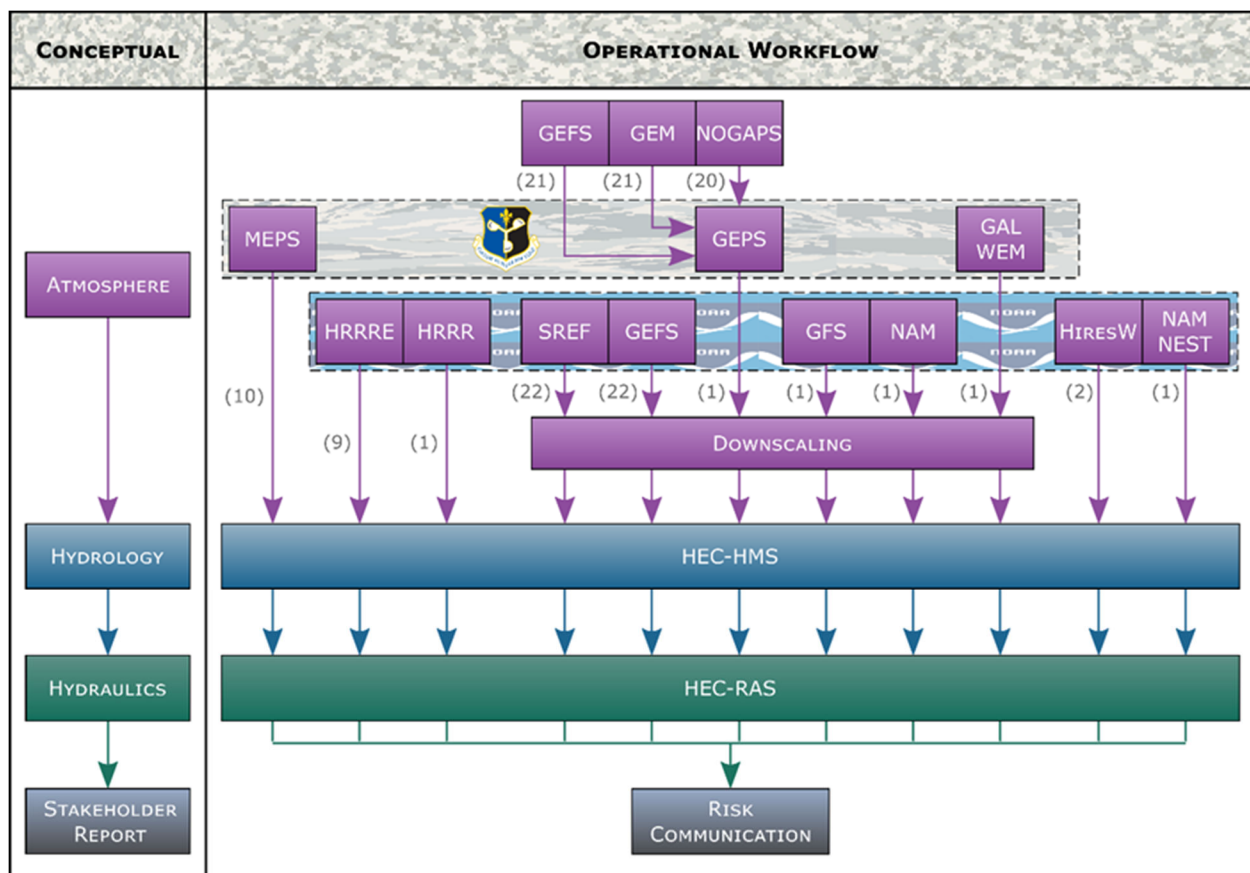


Figure 2.2: Hydro RF conceptual and operational workflow. Numbers in parentheses indicate the number of members in each forecast ensemble and subsequent analyses.

2.1.1 Atmospheric Component

HydroRF considers atmospheric forcing data from six sources, including the High-Resolution Rapid Refresh (HRRR), HRRR Ensemble (HRRRE), Global Forecast System (GFS), North American Mesoscale Forecast System CONUS nest (NAM-CONUS-nest), the HighRes Window (HiResW) products, and the Short Range Ensemble Forecast (SREF) (Table 2.1). While the

HRRR, HRRRE, NAM-CONUS-nest, and HiResW products can be used in their native state, the GFS and SREF datasets must be downscaled to support Hydro RF applications. Information about the different atmospheric datasets and details of the downscaling process are discussed thusly.

HRRR

The High-Resolution Rapid Refresh (HRRR) model is a NOAA real-time 3-km resolution, hourly updated, cloud-resolving, convection-allowing atmospheric model, initialized by 3km grids with 3km radar assimilation over CONUS. Four times a day (00Z, 06Z, 12Z, 18Z) the forecast window extends to 48 hrs (was 36 hrs in 2020-2021 during Hydro RF demonstration period). All other initializations are 18 hr forecasts.

HRRRE

The HRRR ensemble (HRRRE) was an experimental 9-member forecast ensemble at the same 3-km resolution and similar physics as the HRRR. The 9 members were available from 12Z for 24 hr forecasts. The HRRRE was discontinued on 14 July 2021.

NAM-Conus Nest

The North American Mesoscale (NAM) Model is a forecast system run and maintained by NCEP that generates 84-hour 12 km horizontal resolution forecasts for the entirety of North America. The NAM-CONUS-nest is a high-resolution nest initialized by and run in parallel to the parent NAM that produces 60-hour 3 km horizontal resolution forecasts for the CONUS, southern Canada, and northern Mexico. Both versions of the NAM are run using the WRF's Non-hydrostatic Mesoscale Model (NMM) core, i.e. WRF-NMM. NAM-CONUS-nest has a sufficiently high horizontal resolution for capturing convective activity.

HighRes Window

The High Resolution Window, or HiResW, is a dual-core modeling system that aims to provide short-term but high, 2.5 km, horizontal resolution forecasts over the United States and its major territories. The system uses the Advanced Research WRF, known as the ARW core or WRF-ARW, and previously the WRF-NMM core. On 12 May 2021, the WRF-NMM core was replaced with the more modern FV3 model core configured in a limited area capacity to better match the gridded output from WRF than the global output for which it was initially designed.

The WRF-ARW core HiResW is initialized from the RAP and uses the GFS for its lateral boundary conditions, whereas the FV3 core HiResW uses GFS forecasts for both initial and boundary conditions. The dynamics and physics for both of these cores are identical to their parent models, effectively serving as a dynamic downscaling model for the RAP and FV3 over specific regions of interest.

GFS

The operational Global Forecast System (GFS) maintained by the National Center for Environmental Prediction (NCEP) runs four times a day (00Z, 06Z, 12Z, 18Z) and generates 16-day forecasts. The GFS framework incorporates a data assimilation system and forecast models for the atmosphere and ocean waves. The atmospheric part of the GFS uses a Finite Volume Cubed Sphere (FV3) dynamical core. It runs with a horizontal resolution of ~13 km and 127 vertical layers, including levels that extend up into the mesopause (~80 km height). While the system runs at relatively high resolution for a global model, NCEP releases GFS output on a quarter-degree equally spaced grid with 57 vertical standard pressure levels with a temporal resolution of 1 hour out to 120 hours, then 3 hours for days 5-16.

SREF

The Short Range Ensemble Forecast (SREF, Du et al. 2015) is an operational short duration ensemble forecast also maintained by NCEP. The ensemble currently consists of 26 members based on a multi-analysis, multi-model, and multi-physics approach. While the configurations of individual members vary, each produces forecasts with 16 km horizontal resolution and 60 vertical levels over the CONUS up to 87 hours long every six hours (03Z, 09Z, 15Z, 21Z). Members in the SREF are initialized with a three-hour forecast due to its offset from traditional forecast cycle times, from the operational North American Model (NAM), GFS, or Rapid Refresh (RAP) models. Boundary conditions for each are provided by the Global Ensemble Forecasting System (GEFS) and utilize a variety of initial condition perturbations, CU schemes, planetary boundary layer (PBL) schemes, MP schemes, long and short wave radiation schemes, and soil perturbations.

Data from medium-term forecasts (SREF) are too coarse of a spatial and temporal resolution to predict regional precipitation. Downscaling techniques bridge the gap between the resolution of synoptic scale atmospheric models and regional scale processes (Fowler et al., 2007). Thus, Hydro RF addresses coarse model resolution with a built-in WRF capability (e.g. Skamarock et al. 2008; Wei et al. 2009) to dynamically downscale these medium range forecasts to a spatial (1.33 km) and temporal scale (1 hour) sufficient for use in hydrologic models. These resolutions were selected because this is commonly accepted as the spatial resolution where convective processes in atmospheric models will resolve themselves within a single grid cell, and thus do not need to be approximated through a convective parameterization scheme.

Table 2.1: Ensemble member atmospheric forcing data modeling systems and their associated attributes.

Modeling system	Resolution	Members	Forecast length for study	Model cycles			
				00Z	03Z	12Z	15Z
High-Resolution Rapid Refresh (HRRR)	3 km	1	18 hr / 36 hr	Hourly			
HRRR ensemble (HRRRE)*	3 km	9	24 hr			✓	
North American Mesoscale Forecast System CONUS nest (NAM-CONUS nest)	3 km	1	60 hr	✓		✓	
High-Resolution Window (HiResW)	2.5 km	2**	48 hr (arw & nmm) 60 hr (fv3)	✓		✓	
Global Forecast System (GFS)	0.25 deg	1	72 hr	✓		✓	
Short-Range Ensemble Forecast (SREF)	16 km	26	72 hr		✓		✓

* The HRRRE discontinued operation 14 July 2021.

* The HiResW-nmm and HiRes-fv3 datasets were combined into a single member forcing dataset since the HiRes-fv3 replaced the HiResW-nmm product in May 2021.

The precipitation forecasts must be regridded and reprojected to the grid of the hydrology model. Each of the operational forecasts has their own native grid (e.g., Polar Stereographic and Lambert Conformal conic) and they all must be reprojected to the Albers Equal Area projection. An equal area projection is required within the hydrology model to ensure that volumetric calculations are uniform across the grid. The various operational forecasts also all have their own spatial resolutions, therefore they need to be regridding to a uniform spatial resolution for HEC-HMS input. A conservative regridding scheme is used to transform the grids of the precipitation forecasts. The conservative method guarantees that the volume of precipitation is conserved after the transformation. The method is different from other regridding methods, such as nearest neighbor, because maintaining the overall continuity of the entire grid is prioritized over continuity between individual grid cells. An example of the reprojection and regridding is shown in Figure 2.3 for an hour of NCEP Stage IV precipitation over the same subbasin.

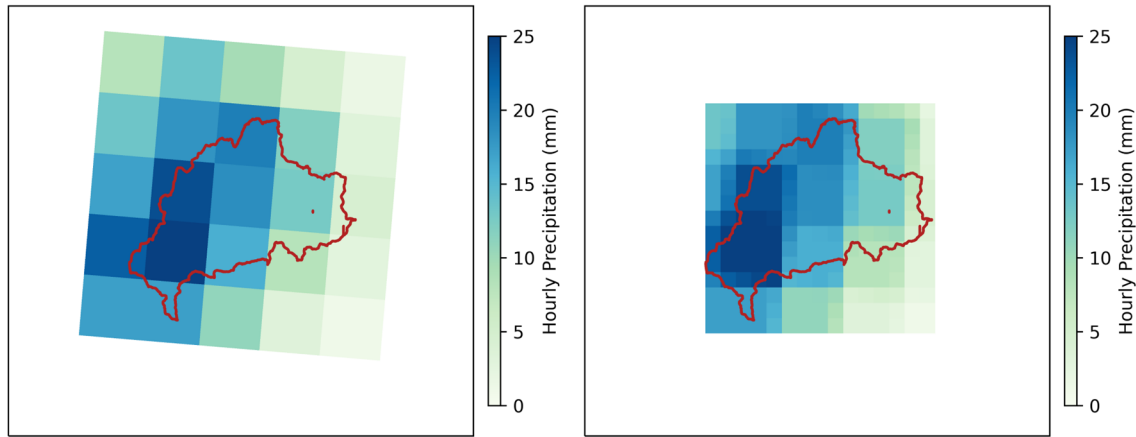


Figure 2.3: An example of the precipitation regridding methodology. Examples of NCEP Stage IV precipitation on its native Polar Stereographic grid (left) and the same NCEP Stage IV data reprojected & regridded (right) to the equal area grid of the HEC-HMS hydrology model.

2.1.2 Hydrology and Hydraulic Component

The surface module is a series of watershed-specific hydrologic and hydraulic models, which maintains a continuous simulation, which updates daily, and forecasts individual streamflow/flooding events and the hydraulic responses at critical infrastructure, using the continuous simulation to hot start the forecast runs. The output of the surface module is the streamflow forecast within a watershed containing the infrastructure of concern. Basin-scale conditions are then further integrated to determine water depth and stream velocity at local points of interest and assess the forecasted usability, safety, and functionality of specific infrastructure. Hydro RF transforms the precipitation grid from each atmospheric ensemble member into excess runoff with the HEC-HMS hydrologic model. Hydro RF requires an underlying hydrologic model at each atmospheric forcing resolution for each specific area of interest, as each model must be calibrated separately per local conditions.

HEC-HMS accomplishes the physical representation of a watershed with a basin-scale model. Widely available GIS data (i.e. topographic, soils, land use, and drainage networks) are used to create a physically based representation of a basin. HEC-HMS accepts gridded, long-term precipitation data which is transformed into saturation and infiltration excess runoff at the gridded sub-basin scale using the gridded Soil Moisture Accounting (SMA) method. The excess runoff is routed overland through the ModClark transform method and then into the channel reaches to the basin outlet using the Muskingum-Cunge method. HEC-HMS models are calibrated using gauge precipitation data or gridded precipitation data (e.g. radar data) in conjunction with stream gauge data to validate modeled hydrographs with observed hydrographs for historic events.

Hydro RF produces hydraulic parameters (e.g. flow depth, flow velocity, top width, water surface elevation) at local points with use of HEC-RAS. HEC-RAS models are developed through field surveys with some additional definition of channel overbank areas and roadway crests using the topographic data. Bridges, culverts, and hydraulic structures are either surveyed or data from as-built plans are used to accurately represent the geometry of the structures within the model. All additional required parameters are determined during surveying. Steady, 1-D flow profiles are

generated for each study reach and available gauge data is used to calibrate each model at each gauge location using flow rate and flow depth data. A wide range of flow profiles are produced in order to develop rating curves (e.g. flow rate vs. flow depth, flow rate vs. flow velocity) at each critical infrastructure location. Runoff generated from HEC-HMS models within Hydro RF are routed through the basin, enabling the prediction of water depth and velocity at an individual spatial point, such as a stream LWC, using the rating curves generated in HEC-RAS (Gambill, 2016).

The output from the hydrology and hydraulic module of Hydro RF consists of time series (equal in number to the number of atmospheric members used) of flow rate at points of interest in a basin. Figure 2.4 shows example hyetographs (blue) and hydrographs (green) for a 2016 event at Fort Hood. The hyetographs show the time series of precipitation averaged across 10 NCAR ensemble members. The hydrographs show the output from HEC-HMS at the USGS gauge for each NCAR ensemble forcing member (light green) and the average streamflow across all 10 members (dark green bold line). USGS stream gauge 08101000 observations are shown in black. Rating curves for points of interest relating flow rate to hydraulic parameters such as flow depth, flow velocity, water surface elevation, etc., will be generated using HEC-RAS.

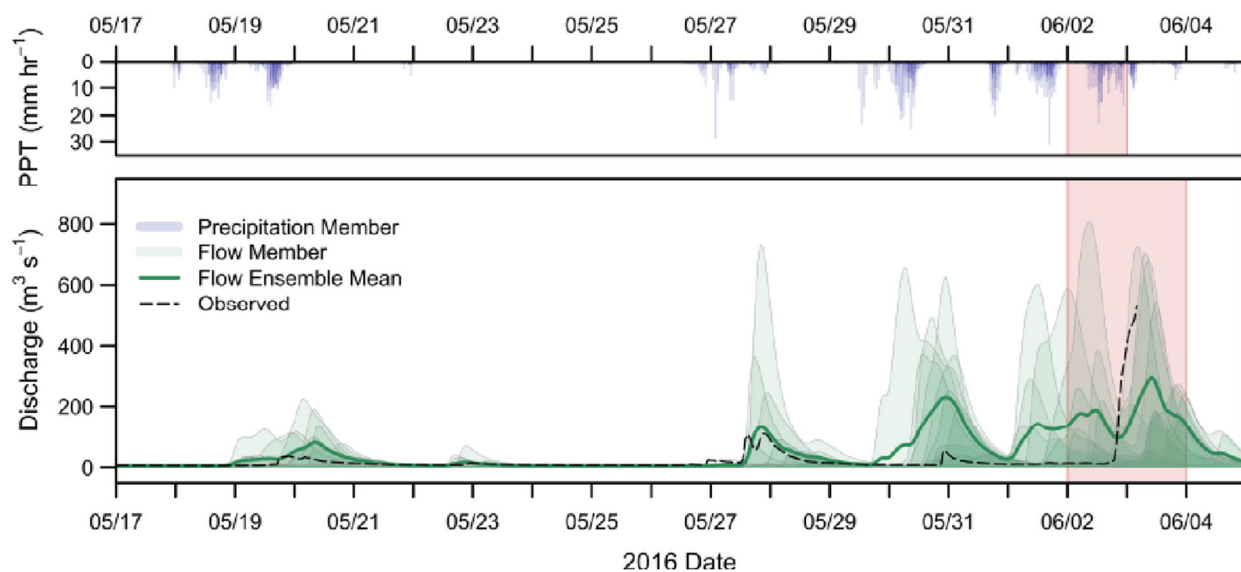


Figure 2.4: Example hyetographs (blue) and hydrographs (green) for a 2016 event at Fort Hood.

2.1.3 Risk Communication Component

Hydro RF uses flow rate time series to generate risk reports at frequencies of 1-hour, 3-hours, 6-hours, 12-hours, 24-hours, and 72-hours. Within each report, Hydro RF isolates if an event is in the forecast and the variance of that event across model members is conveyed in magnitude and arrival time through histograms and probability density functions (PDFs). The magnitude of the event is communicated as a probability of being within prescribed thresholds unique to a site (i.e. water depth and velocity at a LWC).

Hydrologic/hydraulic simulations were compared to high-water categories specified by the NWS. The NWS currently uses high-water level categories at USGS stream gauges to communicate flood hazard. The high-water level terms include bank-full stage, action stage, and flood stage. The action stage is defined by the NWS as, “the stage which, when reached by a rising stream, represents the level where the NWS or a partner/user needs to take some type of mitigation action in preparation for possible significant hydrologic activity” (see www.weather.gov/aprfc/terminology). The flood stage is more severe and defined by the NWS as the water level that is hazardous to life and property.

The high-water levels at USGS gauges aren't applicable at the remote critical infrastructure on military installations. These levels are too generic and don't apply to LWC that are at-grade with the channel bottom. As an alternative to high-water categories, the multiplication of flow velocity and flow depth is herein referred to HST. HST values are set on the stability thresholds that cause vehicles to float, slide, or overturn (Shand et al., 2011). The appearance of HST in literature was reviewed by Martínez-Gomariz et al. (2018), with values from experimental and analytical studies. The thresholds are highly vehicle dependent. A communication strategy that incorporates HST allows tailoring to specific vehicles. Martínez-Gomariz et al. (2018) found that the best to date reference for HST criteria, is from the AR&R guideline (Shand et al., 2011) and used herein. The least conservative AR&R criteria is for stationary, large 4-wheel drive vehicles. These criteria specify HST never exceeds $0.6 \text{ m}^2 \text{ s}^{-1}$ with additional constraints that depth never exceeds 0.5 m and velocity never exceeds 3.0 m s^{-1} (Shand et al., 2011).

2.2 ADVANTAGES AND LIMITATIONS OF THE TECHNOLOGY

Current hydrologic risk management in remote areas depend heavily on in-situ observations to support decisions. One of the most advantageous aspects of observations, from stream gauges and radar for example, is reliability. Decision makers have high confidence that the information being delivered is correct. One of the most disadvantageous aspects of observations though is that they occur in near real-time and constrain the time decision makers have available to prepare and react. A tragic example of how limited decision support can lead to disaster is the training fatalities of nine soldiers at a LWC at Fort Hood on 2 June 2016. The Investigating Officer for the incident states in the executive summary that, “although not directly causing this incident, procedural shortcoming at the installation, company, and platoon levels—specifically, training on local hazards, driver’s training, and LWC hazard reporting—should be improved to more effectively mitigate risks in the future” (DoA, 2016).

Hydro RF transforms current hydrologic risk assessment by incorporating state-of-the-art weather forecasting products and atmospheric, hydrologic and hydraulic modeling. However, it is important to acknowledge the advantages and disadvantages of hydrometeorological forecasting. The longer into the future the prediction is, referred to as the lead time, the greater the uncertainty associated with that prediction. The increase in uncertainty with increased lead time occurs because hydrometeorological systems are chaotic. Chaotic systems are highly sensitive to initial conditions (Lorenz, 1993; Li and Yorke, 1975). Small perturbations in initial conditions (such as differences between reality and observations used to initialize a weather model) will cause the reality trajectory and forecast trajectory to diverge from each other through time, resulting in an inaccurate prediction (Palmer, 2017).

A single forecast trajectory is produced by a deterministic model. Deterministic forecasting methods attempt to produce a single, accurate forecast of an event (e.g. a single answer with associated uncertainty). To account for chaotic system behavior, methodologies exist that combine several of these single model trajectories, but with different initial conditions. The combination of these trajectories is referred to as a probabilistic model, or ensemble. The proposed work minimizes uncertainty by combining high-resolution deterministic predictions with ensemble predictions through probabilistic forecasting. The combination of several model runs can be more telling than any individual model run (referred to as a member). Probabilistic forecasting offers a simple metric to communicate uncertainty to decision makers in a manner that fosters continued trust (e.g. 90% probability of excess rainfall). Open communication of forecast uncertainty is critical for establishing decision maker confidence (Palmer, 2017). For example, if all the ensemble members predict an event at the same future time, but of varying magnitudes, then decision makers can be provided information that covers the realm of possibilities with corresponding probabilities of occurrence. The incorporation of the probability of occurrence within ensemble forecasting prevents decision makers from having to “put all their eggs in one basket” as they would have to with a deterministic forecast which provides a single prediction. A probabilistic approach allows for the uncertainty associated with highly complex environmental interactions to be conveyed in a clear and concise manner. The USAF Weather Agency currently has an operational ensemble addressing uncertainty through the stochastic perturbation of initial conditions and model physics (Hacker et al., 2011; Clements, 2014). However, the current operational ensemble provides only atmospheric variables such as precipitation. The proposed work looks to adapt and expand the USAF’s operational probabilistic precipitation forecast to hydrologic applications for fine scale streamflow prediction.

3.0 PERFORMANCE OBJECTIVES

The performance objectives for the Hydro RF demonstration are provided in Table 3.1. As shown in the table, the performance metrics include both quantitative and qualitative measures. Narrative descriptions are provided in the following subsections, highlighting each objective's relevance to the technology demonstration and validation. Quantitative objectives in the table are based on results from previous studies and modeling experience and are intended to assist in the validation of the model chain for three demonstration sites and stress test of probabilistic ensemble forecasts to maximize efficiency. The qualitative objectives are established for the purpose of facilitating implementation and technology transfer.

Performance objectives 1-3 will be used to validate the model chain for three difference demonstration sites with different characteristics. Hindcasted events (at least one year worth of simulation) will be used to evaluate performance of 1) probabilistic forecasted precipitation from all ensemble members against observed precipitation, 2) probabilistic forecasted streamflow from all ensembles against observed streamflow, and 3) probabilistic forecasted risk from all ensembles against observed risk (using hydraulic model rating curves and observed flow). Performance objective 4 will stress test Hydro RF to determine if/how Hydro RF can be run at less than full capability during active periods while maintaining statistically equivalent forecasted risk reports as those generated by full capability scenarios. Section 5 provides procedures in more detail of how the performance objectives will be utilized within the Test Design to achieve the objectives of the demonstration.

Table 3.1: Demonstration and Validation Performance Objectives and Descriptions.

Performance Objective	Metric	Data Requirements	Success Criteria
Quantitative Performance Objectives			
Accurate probabilistic precipitation forecasts	Fractions Skill Score	● Stage IV data ● Ensemble members ● WRF model(s)	● <24-hr forecast FSS > 0.5 ● >24-hr forecast FSS > 0.2
Good fit between long-term historic and forecasted streamflow time series	Brier Skill Score	● Streamflow gauge data ● Survey Data ● Soil, land use, and DEM ● Calibrated HEC-HMS model	● 72-hr forecast BSS > 0.3
Forecasted risk agreement with historic conditions	Risk assessment forecasted agrees with risk of historic event (Type I/Type II error)	● Same as above ● Calibrated HEC-RAS model and rating curves	● Forecast 75% agreement with obs (True/True and False/False) ● Minimize Type II error
Complete downscaling of ensemble members with good computational efficiency	Brier Score	● Same as above	● WRF-downscaled members are necessary if their associated BS scores are lower than the Ens1 results.
Qualitative Performance Objectives			
Risk Report ease of use	Positive survey results from decision makers	Feedback from decision makers on usability of risk reports	Decision makers able to understand and utilize risk assessment

3.1 QUANTITATIVE PERFORMANCE OBJECTIVES

3.1.1 Accurate probability precipitation forecasts

The grid-scale accuracy of numerical weather prediction (NWP) models run at convection-allowing resolutions (1–5 km) can be severely limited by the rapid growth of small-scale errors (Lorenz 1969), especially in the presence of precipitating systems (e.g., Zhang 2003). It is therefore not uncommon for high-resolution precipitation forecasts to possess some degree of spatial error, even at early lead times and/or when the overall structure of the precipitation system is well simulated. Moreover, a fair assessment of the accuracy of high-resolution forecasts can be hindered by the use of traditional point-based verification metrics, which require modeled and observed values to exactly match in space and time. For example, some measures yield excessively low scores due to the “double penalty” effect in which a forecast offset slightly from the observations will be penalized twice (once as a miss and once as a false alarm) even if the forecast precipitation resides in the correct catchment. Other measures do not account for displacement error at all (e.g., Ahijevych et al. 2009), such that a precipitation object could conceivably receive the same score when placed correctly inside or incorrectly outside a catchment. These deficiencies have motivated the development and use of neighborhood verification methods that allow forecast quality to be assessed across a range of spatial scales (i.e., close forecasts are allowed to receive some credit while gross errors are not).

A neighborhood verification method applied extensively to high-resolution precipitation forecasts is the Fractions Skill Score (FSS; Roberts and Lean 2008), and is used here to verify probabilistic precipitation fields from the atmospheric component of Hydro RF. First, an accumulation threshold is used to create binary probabilities of event occurrence at each grid point in the forecast and observation grids separately. A square neighborhood of user-defined size (e.g., 3×3 or 5×5 grid boxes) is then used to compute the fraction of points in the neighborhood where the threshold is met or exceeded. These fractions, sometimes termed neighborhood probabilities, are then used to compare each point in the forecast grid to its corresponding observation using the FSS.

The generic form of a skill score (SS) is given as:

$$\text{Skill Score} = \frac{A_{for} - A_{ref}}{A_{perf} - A_{ref}}$$

where A is the accuracy measure under consideration, A_{ref} is the accuracy measure for a set of reference forecasts, and A_{perf} is the accuracy measure for a perfect forecast. Skill scores are used to convey the degree of improvement or degradation in forecast quality relative to a benchmark accuracy measure. Modeled after the skill score for the mean squared error (MSE) and the Brier Skill Score (BSS), the FSS is given as:

$$FSS = 1 - \frac{FBS}{FBS_{worst}}$$

where FBS is the fractions Brier score given by:

$$FBS = \frac{1}{N} \sum_{i=1}^N (P_{F(i)} - P_{O(i)})^2$$

where $P_{F(i)}$ is the forecast neighborhood probability and $P_{O(i)}$ is the observed neighborhood probability at the i th grid point in the verification domain. The FBS is akin to the Brier score (Brier 1950) and represents the MSE of the neighborhood probabilities. The FBS differs from the Brier score in that the observed probability $P_{O(i)}$ is not restricted to a dichotomous outcome of either 0 or 1. A perfect forecast will yield a FBS of 0, and poorer forecasts will receive scores closer to 1.

The worst FBS forecast is when all nonzero neighborhood probabilities have no overlap (i.e., the forecast and observed probabilistic fields are entirely offset from one another). This score, FBS_{worst} , is given by:

$$FBS = \frac{1}{N} [\sum_{i=1}^N P_{F(i)}^2 + \sum_{i=1}^N P_{O(i)}^2]$$

and serves as a low-skill reference forecast in the computation of the FSS. The FSS itself takes on values between 0 and 1, with 1 representing a perfect forecast and 0 representing no skill. Note that if no neighborhood is used, the FSS reduces to the Brier skill score with the reference forecast having no skill.

Data Requirements

Data required for this objective include operational NWP model output, WRF model output, and Stage IV precipitation analyses.

Success Criteria

The objective will be considered met if $FSS > 0.5$ for forecasts up to 24 hours lead time and $FSS > 0.2$ for forecasts beyond 24 hours lead time.

3.1.2 Accurate probability hydrologic forecasts

The hydrology validation process compared historic streamflow gauge data with streamflow output from long-term, optimized Hydro RF simulations at various outlooks (lead-times). Several methods for quantifying the goodness-of-fit of observations against model-calculated values have been proposed but none of them is free of limitations and are often ambiguous. When a single indicator is used it may lead to incorrect verification of the model. Instead, a combination of graphical results, absolute value error statistics (i.e. root mean square error, RMSE; RMSE-observations standard deviation ratio, RSR), and normalized goodness-of-fit statistics (i.e. Nash–Sutcliffe Efficiency coefficient, NSE; Kling-Gupta Efficiency, KGE) is currently recommended (Ritter and Muñoz-Carpena, 2013). These statistics will be used for HEC-HMS calibration.

RMSE is a commonly used error index (Singh et al., 2004). RMSE is calculated as:

$$RMSE = \sqrt{\sum_{i=1}^n (O_i - P_i)^2}$$

where O is the i th observation for the constituent being evaluated, P is the i th simulated value for the constituent being evaluated, and n is the total number of observations. Despite the extensive use of RMSE, few have offered guidelines to qualify “satisfactory” performance. Singh et al. (2004) suggested values less than half the standard deviation of the measured data may be considered low for hydrograph comparisons at daily and monthly time steps.

Based on the suggestion of Singh et al. (2004), Moriasi et al. (2007) developed the RSR, which standardizes RMSE using the observations standard deviation:

$$RSR = \frac{RMSE}{\sigma_{obs}}$$

where σ_{obs} is the standard deviation in observations. The optimal value of RSR is zero, which indicates zero RMSE, while RSR can be very large for poorly performing models. The lower the RSR, the better the model performance.

NSE is a normalized statistic that determines the relative magnitude of the residual variance compared to the measured data variance and, as proposed by Nash and Sutcliffe (1970), is defined as one minus the sum of the absolute squared differences between the simulated and observed values normalized by the variance of the observed values during the period under investigation. It is calculated as:

$$NSE = 1 - \left[\frac{\sum_{i=1}^n (O_i - P_i)^2}{\sum_{i=1}^n (O_i - \bar{O})^2} \right]$$

NSE was recommended for use in time-series comparisons by ASCE (1993), Legates and McCabe (1999), Moriasi et al. (2007) and others. NSE is commonly used for hydrograph comparisons, which provides extensive information on reported values. NSE indicates how well the plot of observed versus simulated data fits the 1:1 line. Nash-Sutcliffe efficiencies range from $-\infty$ to 1. $NSE = 1$, corresponds to a perfect match of simulated data to the observed data. Generally, $NSE > 0$ are viewed as acceptable model performance, while $NSE \leq 0$ indicates poor model fit (Moriasi et al., 2007). Santhi et al. (2001) and Moriasi et al. (2007) used $NSE > 0.5$ to evaluate that hydrologic model performance was “satisfactory” for hydrograph comparisons at monthly time steps.

KGE is a statistic increasingly used in conjunction with or replacing NSE. KGE combines the three components of the NSE for model error (correlation, bias, and coefficient of variation) in a different way (Gupta et al., 2009). It is calculated as:

$$KGE = 1 - \sqrt{(r - 1)^2 + \left(\frac{\sigma_{sim}}{\sigma_{obs}} - 1\right)^2 + \left(\frac{\mu_{sim}}{\mu_{obs}} - 1\right)^2}$$

where σ_{obs} is the standard deviation in observations, σ_{sim} the standard deviation in simulations, μ_{sim} the simulation mean, and μ_{obs} the observation mean. KGE values greater than $1 - \sqrt{2} = -0.41$ indicate that a model improves upon the mean flow benchmark, even if the model's KGE value is negative (Knoben et al., (2019).

The primary validation statistics used to evaluate ensemble forecast performance will be BS and BSS.

It is also desirable to explore and summarize the performance of the overall hydrological probabilistic ensemble forecasts in addition to their basic statistical properties. The Brier Score (Brier, 1950) measures the accuracy of probabilistic predictions for discrete events (in this case, hazard or no hazard) and is calculated by:

$$BS = \frac{1}{N} \sum_{t=1}^N (f_t - o_t)^2$$

where BS is the Brier Score, f_i is the ensemble forecasted probability of the flood hazard occurring at time t (f_i ranges from 0 to 1), o_i is the actual outcome of the event at time t ($o_i=1$ if the flood hazard occurred, $o_i=0$ if conditions were not hazardous), and N is the number of forecasts considered. In this form, Brier Scores are equivalent to the ensemble prediction's mean squared error. Resultant values will range from 0 to 1, with a score of 0 indicating a perfect forecast and a score of 1 indicating no skill.

The skill score is calculated as:

$$BSS = 1 - \frac{BS}{BS_{ref}}$$

Where BS_{ref} is a reference or baseline prediction, which we seek to improve upon. When assessing streamflow or related variables the reference BS is often the daily average streamflow at a location for the given day over many years. For this project, the gauge data available at most locations is very limited, preventing a traditional climatological reference. Instead, we assume persistent conditions throughout each forecast window by assuming the observed flow (hazard) at the beginning of each forecast remains unchanged for the duration of the forecast window. In this format, we test if the ensemble forecast is skillful over the assumption that present conditions are always persistent.

Data Requirements

Data required for this objective include functional WRF model(s) and input ensemble members. HEC-HMS models and GIS data needed to build the models (i.e. soil, ground cover, and topographic data) are required, as well as streamflow gauge data to streamflow and rainfall gauge data for calibration.

Success Criteria

The objective will be considered met if $BSS > 0.3$ for the entire 72 hr forecast window at most locations.

3.1.3 Forecasted risk agreement with historic conditions

Preliminary risk assessment reports will be generated for each demonstrational infrastructure. Within each report, Hydro RF isolates if an event is in the forecast and the variance of that event across model members is conveyed in magnitude and arrival time through probability density functions (PDFs). The magnitude of the event is communicated as a probability of being within prescribed thresholds unique to a site (i.e. water depth or water surface elevation at a pumping station). The risk assessments provided by Hydro RF forecasts will be compared to historic thresholds as determined by gauge data for corresponding events.

Table 4 shows the Type I/Type II error for safety threshold forecasts. The Type II error, the false negative, occurs when Hydro RF fails to forecast an exceedance when an exceedance is reached. The Type I error, the false positive, occurs when Hydro RF forecasts an exceedance when the safety threshold is not met. The risk assessments provided by Hydro RF forecasts will be compared to historic thresholds as determined by gauge data for corresponding events.

Figure 3.1: Type I/Type II Error Table

Contingency Table Type I/Type II Error Table		Event Observed	
		Safety Threshold Not Reached	Safety Threshold Exceeded
Event Forecast	Safety Threshold Not Reached	True Negative	False Negative (type II error)
	Safety Threshold Exceeded	False Positive (type I error)	True Positive

Data Requirements

Data required for this objective is the same as objective 3.1.2. Plus calibrated HEC-RAS model.

Success Criteria

The objective will be considered met if at least 75% of hazard forecast timesteps agree with observed hazard (true negative/true positive) and type II error is or can be minimized via calibration.

3.1.4 Downscaling of ensemble members with good computational efficiency

We used Brier Score to objectively evaluate the performance of different ensemble configurations considered for the downscaling stress test. The Ensemble 1 (Ens1) configuration represents the degree of accuracy a forecaster or decision-maker could hope to achieve with the probabilistic HydroRF capability over a 48-hour forecast period using limited computational resources (only high-resolution forecast members). In theory, the Ens1 configuration could be run entirely on a laptop or desktop machine without the use of supercomputing resources. Accordingly, we used the Ens1 configuration as our baseline for comparison. The Brier Score measures the accuracy of probabilistic predictions for discrete events (in this case, hazard or no hazard). Brier Scores are equivalent to the ensemble prediction's mean squared error. Resultant values range from 0 to 1, with a score of 0 indicating a perfect forecast and a score of 1 indicating no skill.

Data Requirements

Data required for this objective is the same as objective 3.1.2.

Success Criteria

We consider the WRF-downscaled members necessary if their associated Brier Skill scores are lower than the Ens1 results.

3.2 QUALITATIVE PERFORMANCE OBJECTIVES

3.2.1 Risk Report ease of use

The goal of the proposed work is to demonstrate and validate a tool for Range Control, DPW, DES, or other decision makers that uses near-term forecasts to limit risk to soldiers by optimizing the location and timing of training and limit the risk of infrastructure failure. By ingesting the probability of forecasted risks, decision makers can work to lower the risks or mitigate the corresponding fallout. However, risk reporting is only effective if decision marker can easily understand and test the report, appreciate the probable risks, and act accordingly to mitigate the risks. As a result, preliminary risk assessment reports will be provided to installation decision makers so that they can provide feedback to report improvements. This task will involve two-way conversations between developers and users through feedback/survey/meetings.

Data Requirements

Data from decision makers will be collected to assess the clarity and usability of the Risk Reports and determine the level of overall confidence and the actionability (willingness-to-act) of the decision makers based on the Risk Report.

Success Criteria

The objective will be considered to be met if a majority of decision makers are able to understand, appreciate, and utilize the risk assessments and are willing to take mitigating actions.

4.0 SITE DESCRIPTION

4.1 FORT HOOD

4.1.1 Location and History

Fort Hood, centrally located near Killeen, Texas, is the headquarters of the 3rd Corps. The installation encompasses 890 sq. km. The majority of the installation drains to Belton Lake to the east via Owl Creek and Cowhouse Creek and its tributaries, Table Rock Creek and House Creek (Figure 4.1). These stream channels are typically deeply cut compared to the surrounding land surface, with steep sidewalls and bedrock bottoms. Numerous north-south roads bisect the river channels, which cross via bridges or numerous LWCs.

The Cowhouse Creek watershed is contained primarily within Mill, Hamilton, Coryell Counties with small portions in Comanche, Lampasas and Bell Counties. Upstream of Fort Hood the basin covers 1180 sq. km and a total of 1798 sq. km just upstream of Belton Lake. Table Rock Creek and House Creek basins, tributaries to Cowhouse Creek, cover 168 sq. km and 172 sq. km, respectively. Owl Creek rises within the boundary of Fort Hood and is contained within Coryell and Bell Counties. The basin covers 98 sq. km.

The LWCs on Fort Hood have been a topic of discussion since the deadly 2016 flash flood on Owl Creek and a second deadly flash flood in 2017. Fort Hood DPW has replaced multiple non-tactical, vented LWCs with bridges to improve safety for POVs. However, the tactical LWCs are necessary for training operations on Fort Hood. Without them, heavy forces would not be able to traverse the training areas and execute their training missions properly. LWC replacement with bridges is impractical due to the required weight limits, reduced opportunity for stream fording training, and the large number of existing crossings.

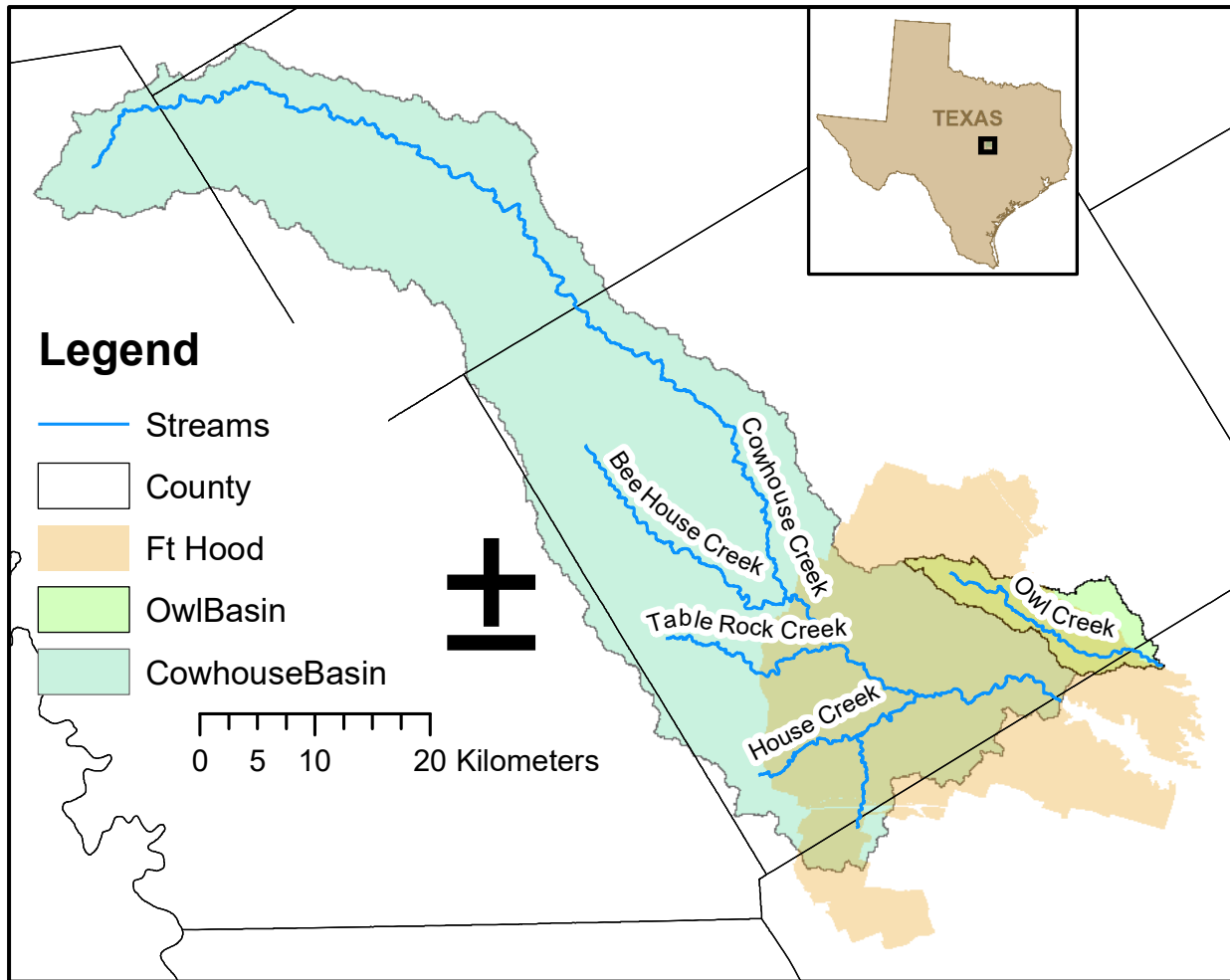


Figure 4.1: Location of Fort Hood within Texas and of the demonstration watersheds (Cowhouse and Owl Creeks) with regards to the installation boundary.

4.1.2 Site Characteristics

The demonstration at Fort Hood is located in the heavy forces maneuver/training areas of the installation within the Cowhouse Creek and Owl Creek watersheds. The training areas have maneuver trails and tactical LWCs where the trails intersect rivers. The demonstration sites are tactical LWCs that are improved fords or vented crossings (Figure 4.2). Improved ford LWCs are at-grade or slightly elevated with respect to the creek bed, requiring all flow to pass over the driving surface. Vented crossing include small culvers which allow baseflow and very frequent events to pass under the driving surface, but which are not large enough to prevent frequent flooding of the driving surface. Such infrastructure types are at the greatest risk to flood hazards.



Figure 4.2: A typical improved ford LWC (FH21, left) and vented LWC (FH01, right) at Fort Hood.

The topography of the demonstration basins is gradually sloping terrain over the floodplains, with steep slopes along the banks of the stream network itself. This topography creates a stream network that is deeply incised (Figure 4.3). The geometry of the channel cross sections is dangerous because the steep banks prevent individuals from being able to escape the flood waters. The land cover characteristics describe a catchment that is undeveloped and remote. The catchment is mostly open range land used for military training and grazing cattle. The 2016 NLCD has 65% of grid cells classified as grassland or shrub and 10% classified as developed. The infrastructure along Cowhouse Creek consists primarily of stream crossing structures, with a lack of flood mitigation infrastructure, such as dams. Three USGS streamflow gauges are located on the main stem of Cowhouse Creek and two on House Creek, all within Fort Hood (Figure 4.4). Three HOBO water level data loggers were installed on Table Rock Creek (2) and Owl Creek (1) to provide observed flow data during the demonstration period. Only one of the USGS gauges, Pidcoke on Cowhouse Creek near the western installation boundary, has been in operation since before 2018.

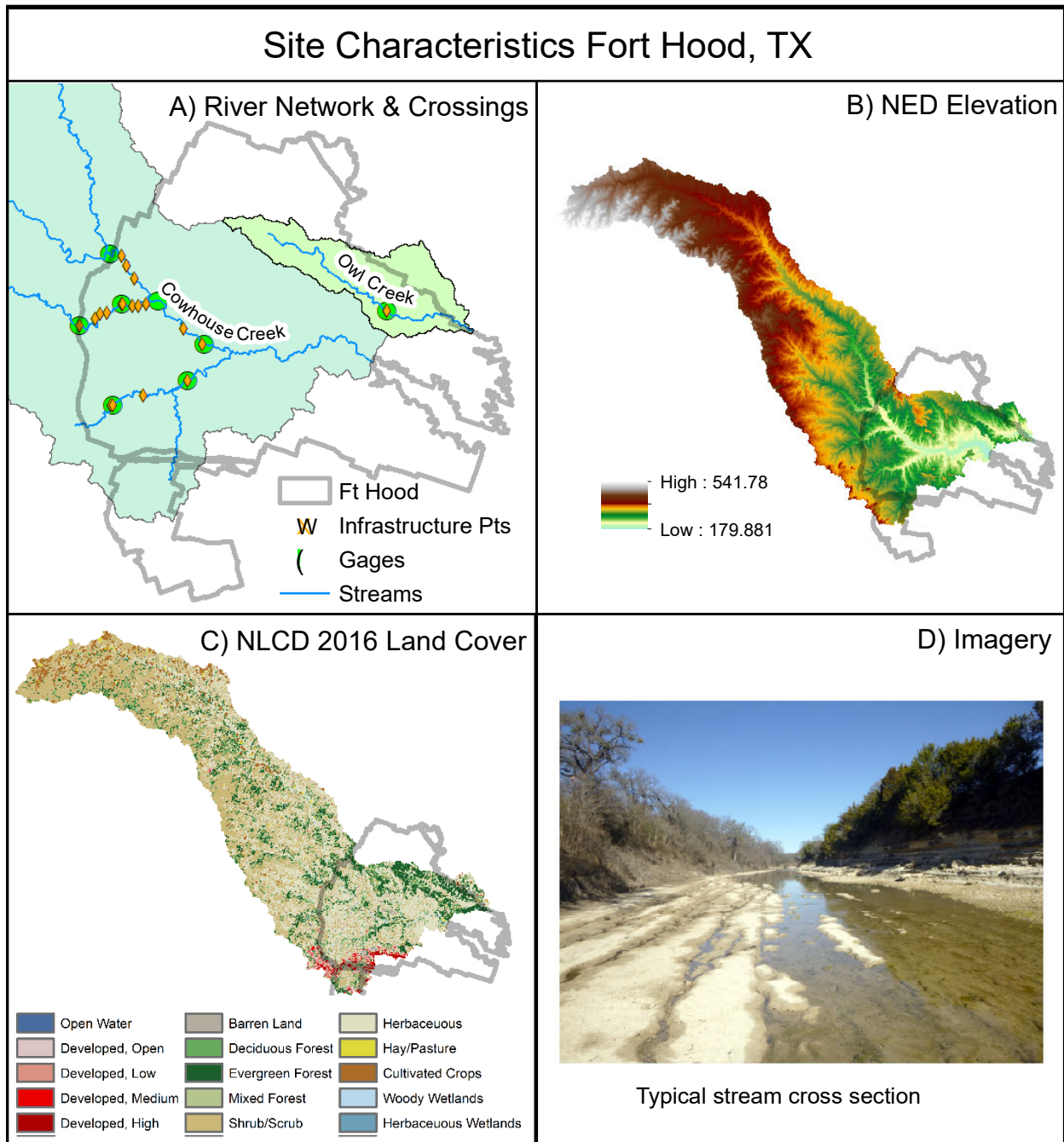


Figure 4.3: Physical characteristics for Cowhouse and Owl Watersheds. A) The stream network for Cowhouse Creek and Owl Creek along with the observation gauges and demonstration infrastructure. B) The NED topography of the watershed. C) The NLCD land cover classifications. D) A typical demonstration reach cross section.

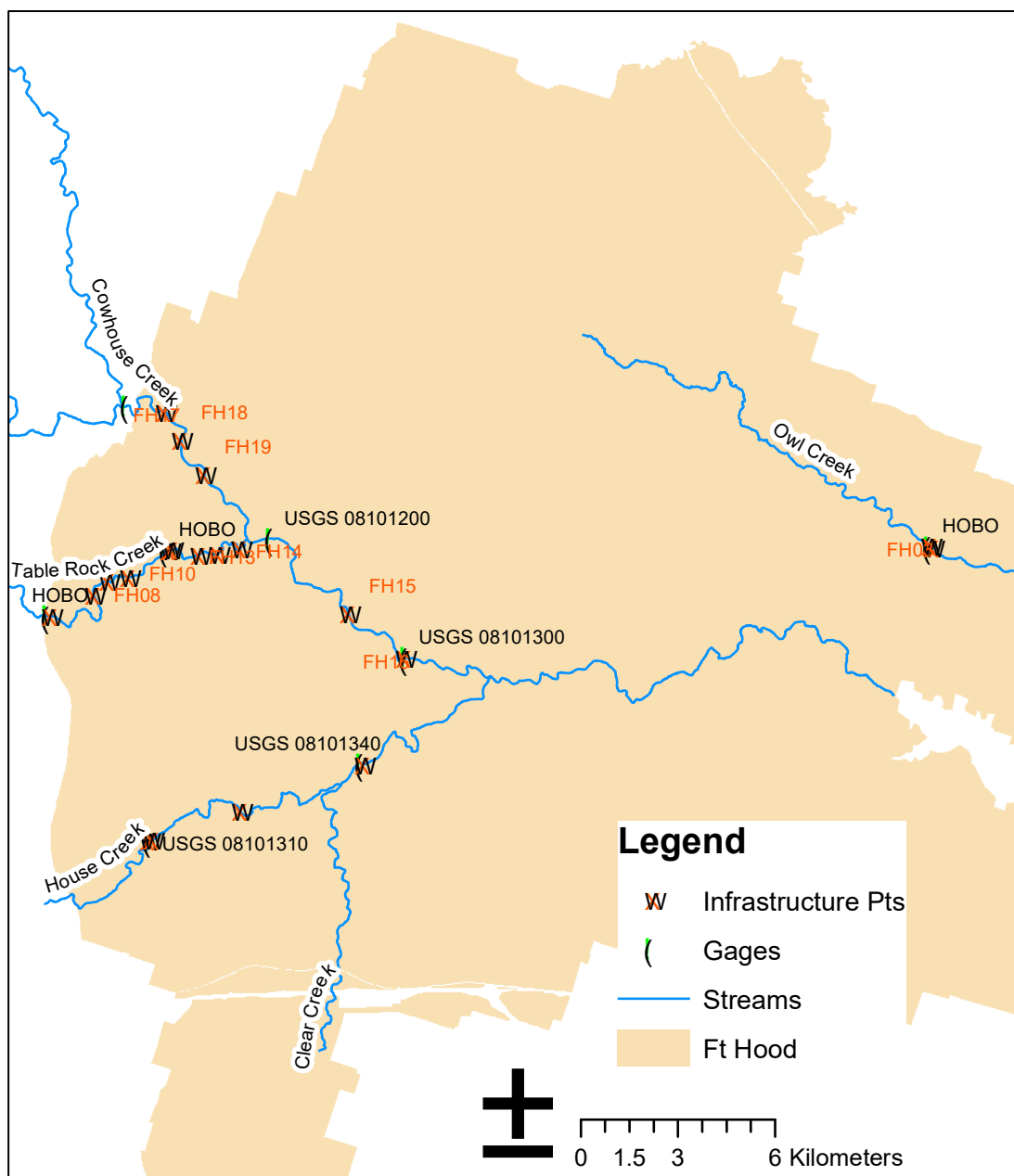


Figure 4.4: Detailed map of stream networks for Cowhouse Creek and Owl Creek along with the observation gauges and demonstration infrastructure.

4.2 FORT LEONARD WOOD

4.2.1 Location and History

Fort Leonard Wood (FLW), located near St. Robert in south central Missouri, is home to the Army Engineer School and Transportation Corps. The installation encompasses 253 sq. km within the Salem Plateau of the Ozarks. The installation drains to the north via Roubidoux Creek and its tributaries Turnbull Hollow and Ballard Hollow, and the Big Piney River (Figure 4.5). These

stream channels are typically shallow and wide with bedrock bottoms while the floodplains are confined within narrow valleys. Numerous installation roads bisect Roubidoux Creek and its tributaries over improved fords and vented LWCs. Significant infrastructure is located within the Big Piney floodplain, including a number of training areas and ranges and the potable water intake pumping station for FLW. A bridge that crosses the Big Piney accesses the east gate of FLW.

The Roubidoux watershed is contained primarily within Pulaski and Texas Counties with small portions in Laclede and Wright Counties. Upstream of FLW the basin covers 460 sq. km and a total of 713 sq. km just downstream of the installation. The Turnbull Hollow basin, a tributary to Roubidoux Creek, covers 28.6 sq. km. The Big Piney watershed is contained within Pulaski, Texas, and Phelps Counties and a small portion of Howell County. The basin covers 1420 sq. km upstream of FLW and 1507 sq. km just downstream.

The May 2016 flood at FLW, which disrupted the potable water supply and restricted access through the East Gate, demonstrated installation vulnerabilities. Training land access is also periodically restricted west of Roubidoux Creek when LWCs are inundated. A POV was swept downstream after attempting to traverse one of the Roubidoux LWCs during flood conditions in the spring of 2019. Flash flooding forewarning can limit the safety risk to personnel and training mission interruptions by optimizing the location and timing of training, while the risk of infrastructure failure can be limited with preemptive mitigation efforts.

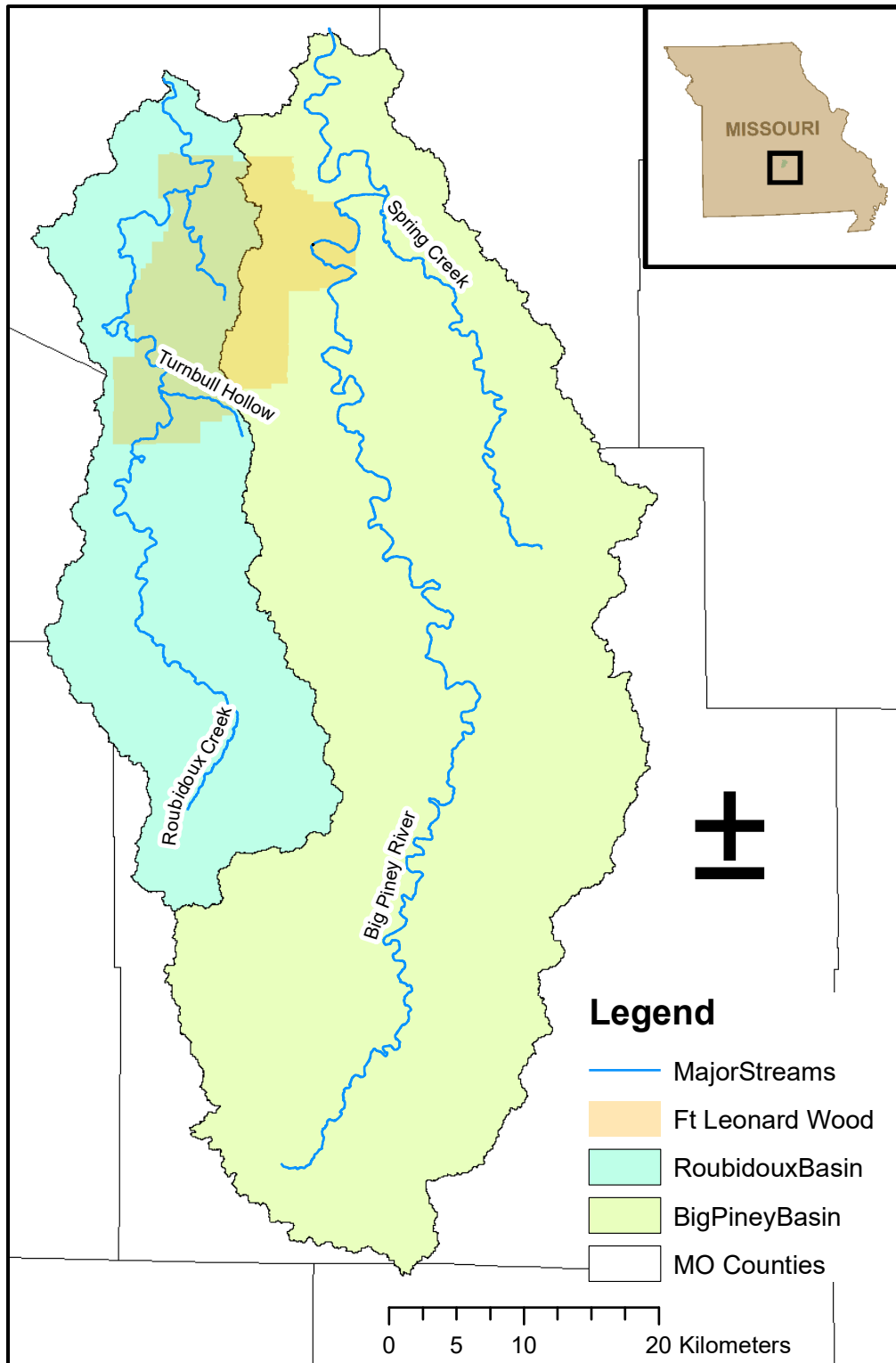


Figure 4.5: Location of Fort Leonard Wood within Missouri and of the demonstration watersheds (Roubidoux Creek and Big Piney River) with regards to the installation boundary.

4.2.2 Site Characteristics

The demonstration at FLW includes three types of riverine infrastructure. Like the other installations, there are LWCs along Rubidoux Creek within the training areas of FLW and also some public LWCs just west of the installation. There is also training range infrastructure susceptible to flash flooding along the Rubidoux and the Big Piney. The potable water intake for the installation with a nearby low head dam and the East Gate access are located along the Big Piney River (Figure 4.6).



Figure 4.6: A typical improved ford LWC (FLW03, top left) and vented LWC (FLW02, top right) on Roubidoux Creek. Low head dam and potable water intake on the Big Piney River (FLW13, bottom left). Typical vented LWC (FLW04, bottom right) on Turnbull Hollow.

The demonstration basins, located in the Salem Plateau of the Ozarks, are characterized by gradually sloping uplands and deep, steeply sloped valleys (hollows) cut by tributaries (Figure 4.7). Karst topography dominates the area due to limestone and dolomite bedrock, with springs and sinkholes common. The karst topography greatly affects the baseflow conditions of the study streams. Roubidoux Creek is a losing stream in some sections of the study reach and a gaining stream in other sections. During base flow conditions, Roubidoux Creek alternates between completely dry reach sections and low flow reach sections, as the flow enters the near surface groundwater at some points and then reemerges into the channel at springs. The Big Piney

maintains baseflow along the study reach, fed by numerous springs. The stream channels are typically shallow and wide with bedrock bottoms while the floodplains are confined within narrow valleys. Agricultural pasture land cover dominates the upper half of the Roubidoux Basin and the upper third of the Big Piney Basin while the remainder of both basins are predominantly deciduous forest. Pastureland constitutes 28% of the study basins while forest covers 56%. The only significant developed area is the cantonment area of FLW.

The infrastructure along Roubidoux Creek consists primarily of stream crossing structures and several training ranges while infrastructure along the Big Piney River consists of training ranges, the East Gate Bridge, and the pumping station (Figure 4.8). Two USGS streamflow gauges are located on each stream, one upstream and one downstream of the installation (total of four). Two HOBO water level data logger were installed on Turnbull Hollow (1) and Roubidoux Creek (1) to provide additional observed flow data locations during the demonstration period.

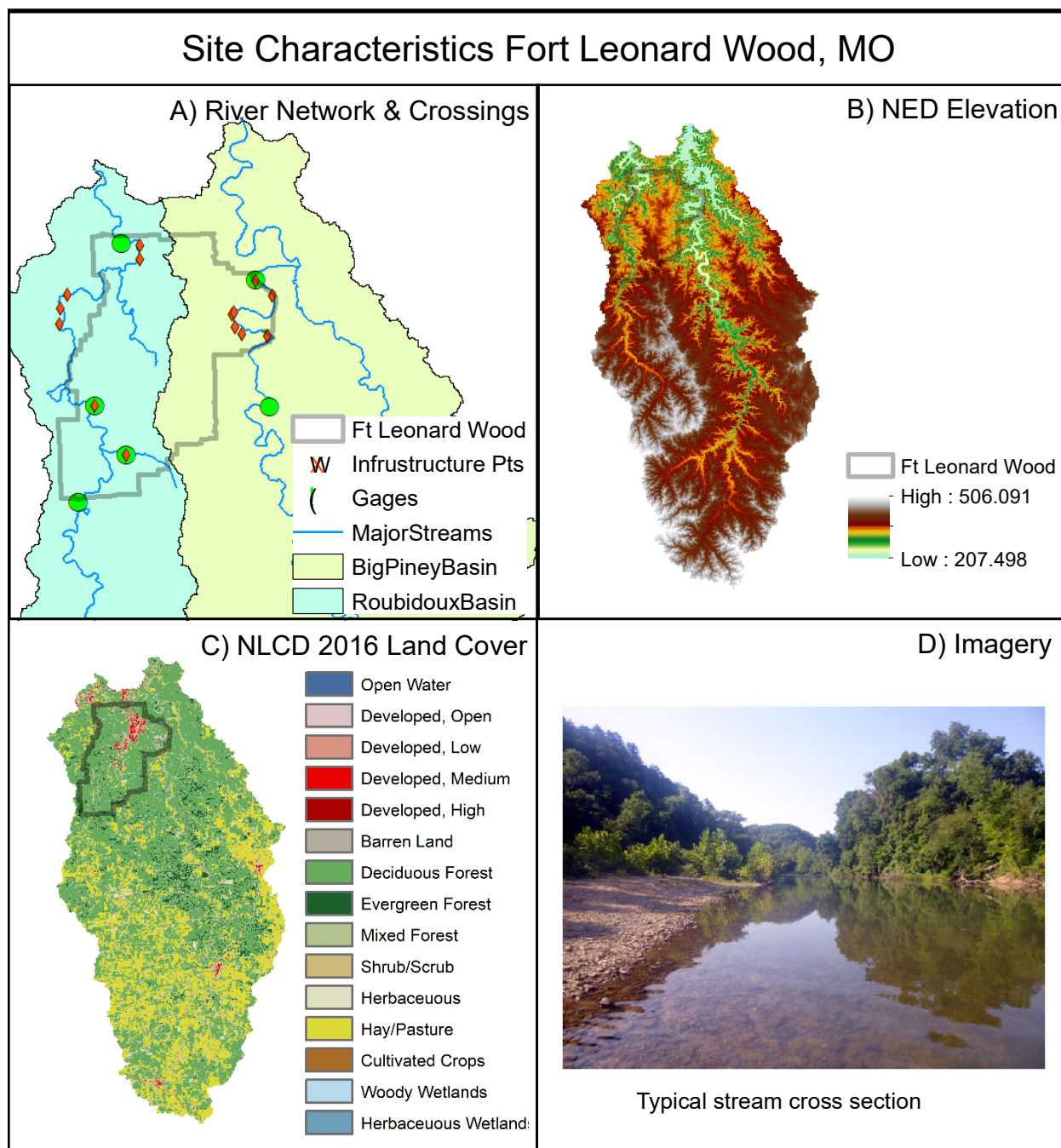


Figure 4.7: Physical characteristics for the hydrologic watershed that contains the demonstration sites at Fort Leonard Wood. A) The stream networks for Roubidoux Creek and Big Piney River along with the observation gauges and demonstration infrastructure. B) The NED topography of the watershed. C) The NLCD land cover classifications. D) A typical demonstration reach cross section.

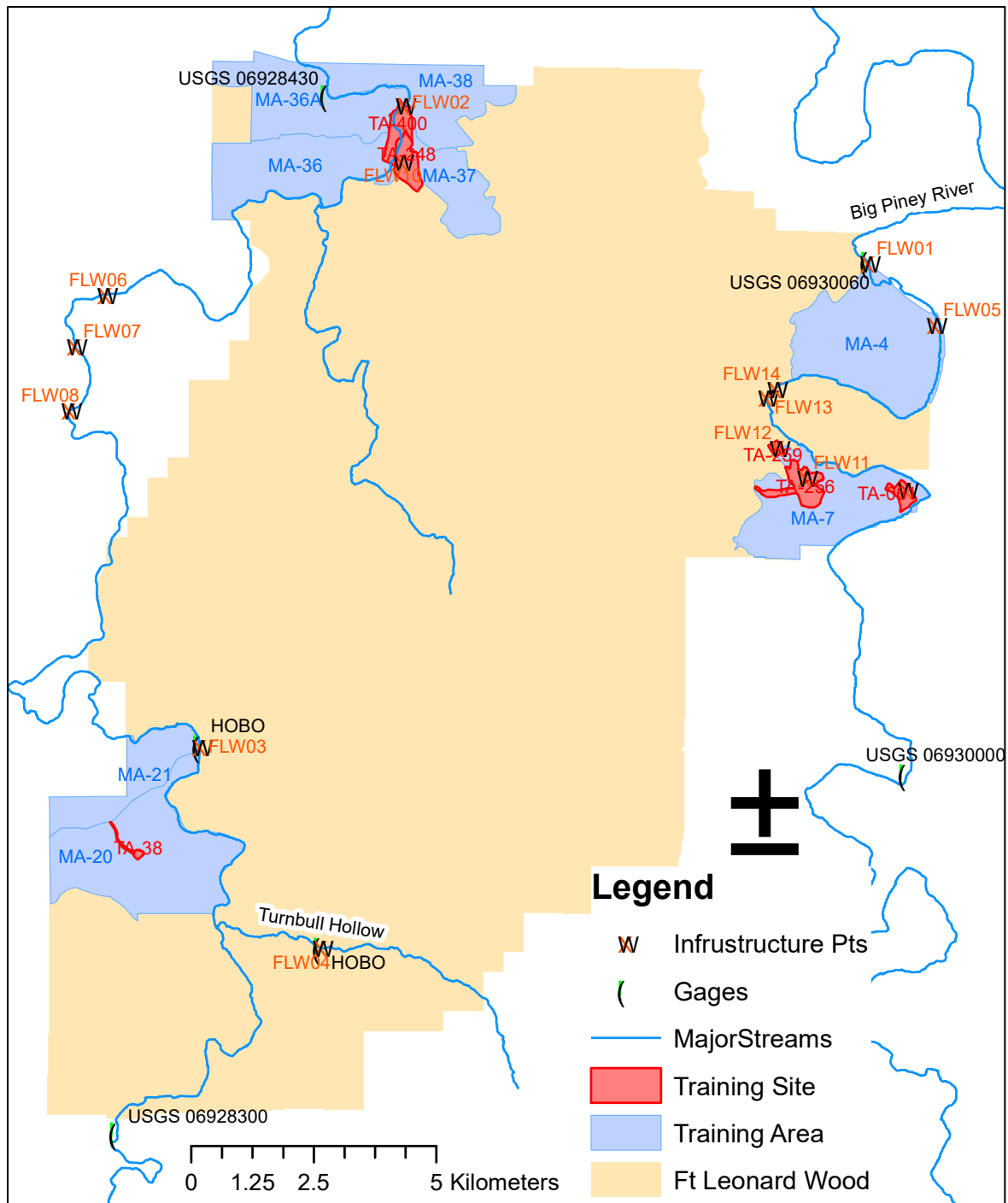


Figure 4.8: Detailed map of stream networks for Rubidoux Creek and Big Piney River along with the observation gauges and demonstration infrastructure.

4.3 FORT CARSON

4.3.1 Location and History

Fort Carson, located just south of Colorado Springs in central Colorado, is home to the 4th Division and 10th Special Forces Group among other units. The installation encompasses 558 sq. km along the eastern base of the front range of the Rocky Mountains.

A majority of the installation drains to the east via tributaries to Fountain Creek while a portion drains to the south via Red and Turkey Creeks, tributaries to the Arkansas River. The demonstration stream for this project is Little Fountain Creek, a tributary of Fountain Creek. (Figure 4.9). The stream channels are typically narrow and steeply sloped. Numerous installation roads and tank trails bisect Little Fountain Creek, which runs west to east, over improved ford and vented LWCs. The cantonment area is located north of the Little Fountain Basin and numerous heavy maneuver training areas are located south of the basin, meaning impassable crossings within the Little Fountain Creek basin can prevent access to the majority of Fort Carson's training lands.

The Little Fountain watershed is contained primarily within El Paso County with only a small portion of the headwaters within Teller County. Upstream of Fort Carson the basin covers 44.4 sq. km and a total of 73.2 sq. km just downstream of the installation.

Little Fountain is not as large as the other creeks and rivers already discussed for the other demonstration sites and flows are intermittent, usually occurring due to convective storms occurring over the Front Range. Fort Carson LWCs flood less frequently compared to the other sites but the floods tend to be more flashy and quick moving due to the basin climate and topography.

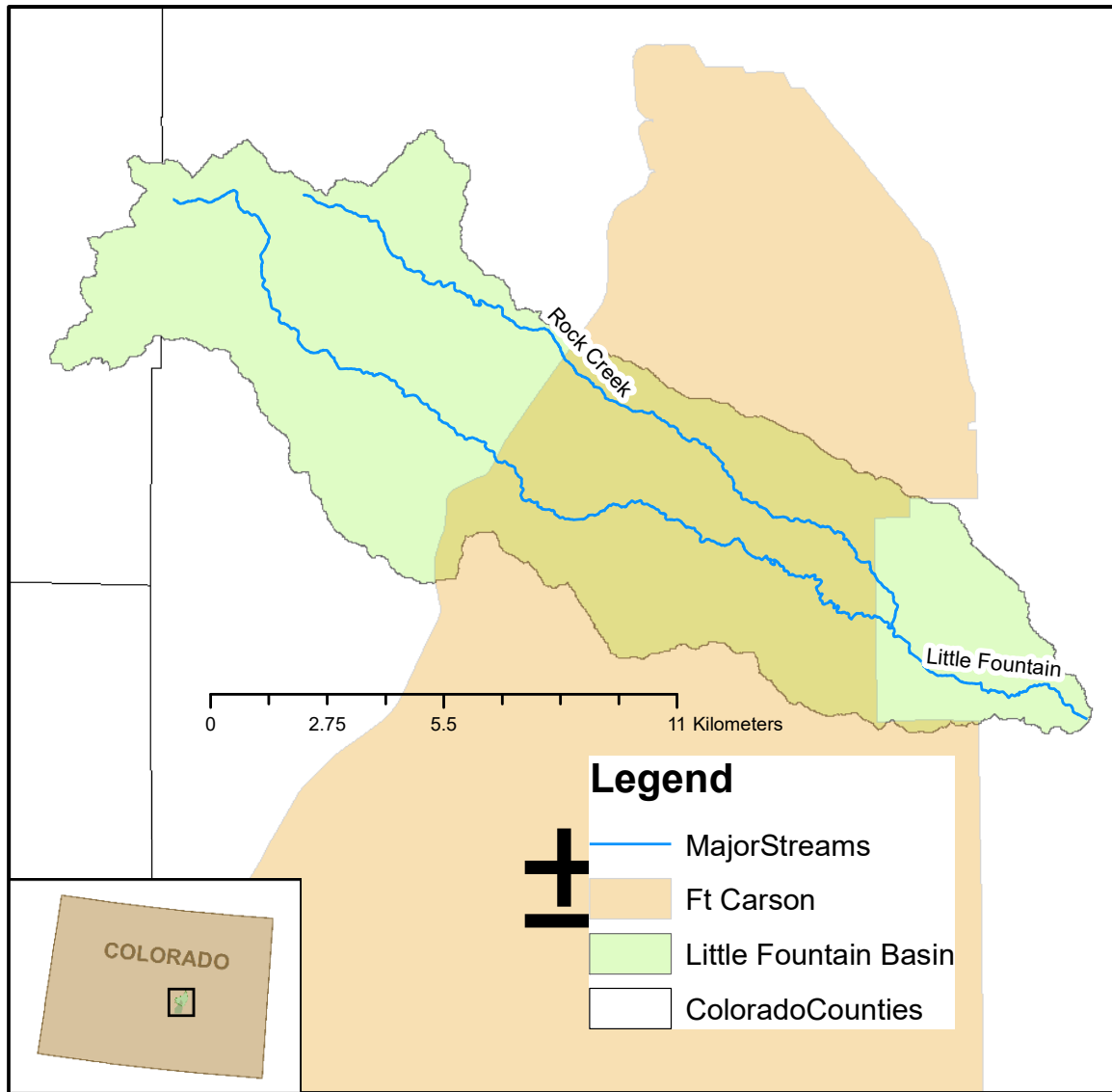


Figure 4.9: Location of Fort Carson within Colorado and of the demonstration watershed (Little Fountain Creek) with regards to the installation boundary.

4.3.2 Site Characteristics

The demonstration at Fort Carson is located in the heavy forces maneuver/training areas of the installation within the Little Fountain watershed. Four tank trails and 3 improved roads cross Little Fountain Creek and provide access to maneuver/training areas to the south. The demonstration sites are four tactical LWCs and one culvert structure (Figure 4.10).



Figure 4.10: A typical improved ford LWC (FC04, left) and vented LWC (FC05, right) at Fort Carson.

The topography of this watershed is unique. The western portion of the watershed is the Rocky Mountains, including Knights Peak at an elevation over 3190 m (Figure 4.11). The demonstration infrastructure is located at elevations between 1700 m and 1850 m. The basin profile is steep (~1300-1500 m drop in elevation) and streamflow energy is high through the stream channels. The land cover is also variable and primarily determined by elevation. The western portion upstream of Fort Carson is in the mountains above 2000 m and predominantly covered by evergreen forest. The downstream two-thirds of the basin is mostly open range shrub/scrub/herbaceous groundcover with herbaceous wetlands near the main drainage channels. The wetland areas have no standing water unless the streams are actively flowing, but trees are more common in these areas. Both Little Fountain Creek and its tributary, Rock Creek, have a USGS gauge with historical data upstream and/or downstream of the infrastructure selected for the demonstration.

The infrastructure along Little Fountain Creek consists primarily of stream crossing structures (Figure 4.12). One USGS streamflow gauge is located just upstream of the confluence with Rock Creek.

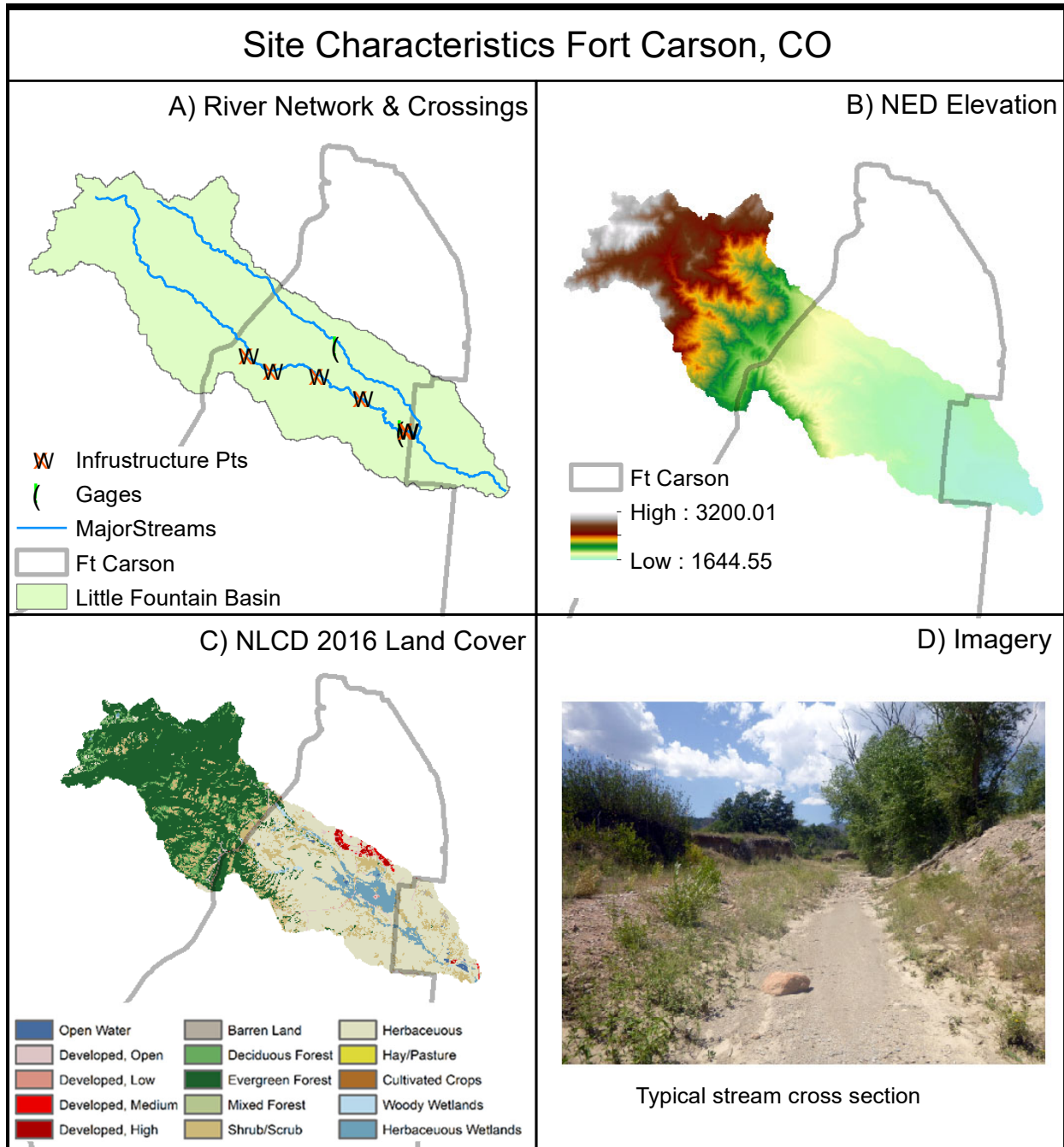


Figure 4.11: Physical characteristics for the hydrologic watershed that contains the demonstration sites at Fort Carson. A) The stream network for Fountain Creek and its tributaries along with the observation gauges and maneuver trail crossings. B) The NED topography of the watershed. C) The NLCD land cover classifications. D) A typical demonstration reach cross section.

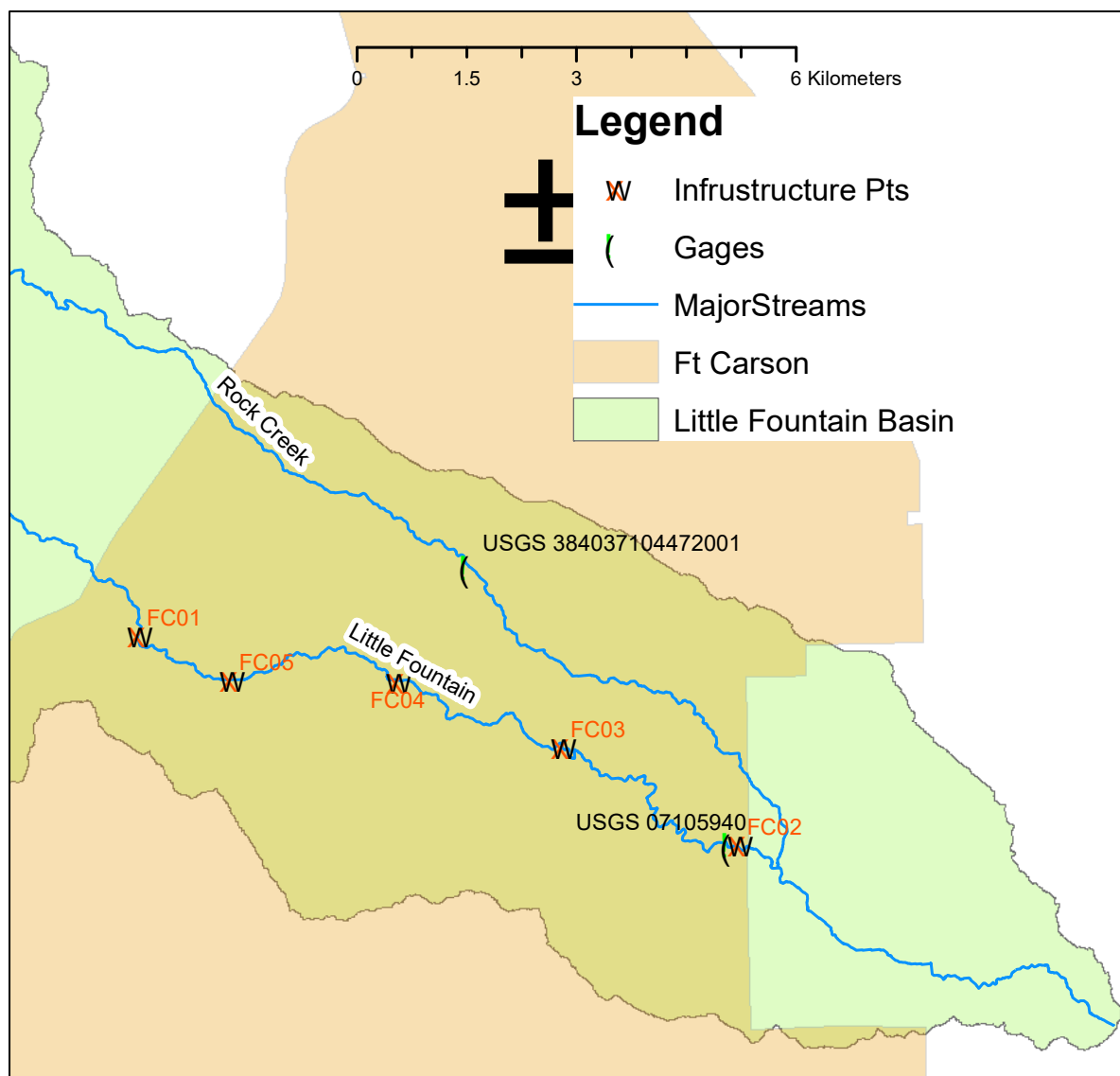


Figure 4.12: Detailed map of stream network for Little Fountain Creek along with the observation gauges and demonstration infrastructure.

5.0 TEST DESIGN

The purpose of Hydro RF is to operationally predict probabilistic flooding hazard at critical infrastructure on military installations through the ingestion and ensemble of weather forecasts.

To demonstrate and validate Hydro RF, the following performance objectives must be met:

1. Accurate probabilistic precipitation forecasts
2. Good fit between long-term historic and forecasted streamflow time series
3. Forecasted risk agreement with historic conditions
4. Complete downscaling of ensemble members with good computational efficiency
5. Risk Report ease of use

Therefore, the test design of Hydro RF will be conducted as follows to address these performance objectives.

5.1 CONCEPTUAL TEST DESIGN

Hydro RF is a model chain that ingests regional weather forecasts and outputs hydrologic risk reports at specific points of interest. There are several steps in this chain that occur “under the hood” of Hydro RF and away from users’ eyes. Performance objectives 1–3 are hard checks on Hydro RF while the capability test was conducted to guarantee that the sub-modules of Hydro RF are performing well.

The greatest utility of Hydro RF is when there is severe weather in the forecast and the predictions of that weather are uncertain. Performance objectives 1-3 are intended to be met leading up to and during active periods of Hydro RF for a successful test of this technology. Hindcast forecast member simulations are run for 2020 (01 April - 13 June) and 2021 (01 March - 15 July) evaluation periods to produce probabilistic forecasts for precipitation, streamflow, and risk. The forecasts are compared to the observed data using the performance assessments. This process constitutes model validation of Hydro RF at three demonstration locations.

Performance objective 4 of the test design aims to stress test Hydro RF to determine if and how Hydro RF can be run efficiently, either by reducing the number of forecast members or assessing the utility of dynamically downscaled forecast members. Hydro RF uses the WRF model to downscale coarser-resolution atmospheric forcing datasets to obtain sufficiently high-resolution mesoscale forecasts of precipitation with forecast windows that extend further than high resolution operational forecasts. This performance objective is tested by running Hydro RF at all three demonstration installations using a variety of capability scenarios. Statistical differences are tested using performance objectives 1-3.

The last performance objective is intended to make Hydro RF as useful as possible to those that will eventually use it. If users don’t understand the reports, then it doesn’t really matter how well Hydro RF performs. The test uses two-way communication to introduce prototype risk report communication strategies to potential users, receiving feedback, and revising the prototypes based on the feedback.

5.2 BASELINE CHARACTERIZATION AND PREPARATION

Surveys were completed to capture detailed infrastructure data with some additional definition of channel geometry, channel overbank areas and roadway crests. Bridges, culverts, hydraulic structures, and all other demonstration infrastructure were either surveyed, measured by hand as appropriate, or data from as-built plans were used to accurately represent the geometry of the structures within the HEC-RAS model. All additional required parameters were determined during surveying and current conditions of the watershed, stream channel, overbank areas, and infrastructure were recorded with field notes and pictures. LiDAR data was used to supplement survey data while soil and landcover data sets were used to supplement parameter determinations made during site visits.

5.3 DESIGN AND LAYOUT OF TECHNOLOGY AND METHODOLOGY COMPONENTS

5.3.1 Water Depth Sensors

HOBO water level data loggers were installed within subbasins lacking USGS stream gauge stations to provide continuous monitoring during 2020 and 2021 (Figure 5.1). The HOBO 30-Foot Depth Water Level Data Logger (Part # U20-001-01) was utilized. The logger is factory calibrated and designed for continuous use. Survey data was used at logger locations to develop rating curves, which were then be used with the logged water depth data to generate flow hydrographs, using the same methods used at USGS gauge sites. HOBO deployment was delayed at FLW due to Covid19 travel restrictions at the beginning of 2020.



Figure 5.1: HOBO water depth data loggers deployed on Table Rock Creek at Fort Hood, TX and on Turnbull Hollow at FLW, MO.

5.3.2 Atmospheric Component

FSS is used here to verify probabilistic precipitation fields from the atmospheric component of Hydro RF. First, an accumulation threshold is used to create binary probabilities of event occurrence at each grid point in the forecast and observation grids separately. A square

neighborhood of user-defined size (e.g., 3×3 or 5×5 grid boxes) is then used to compute the fraction of points in the neighborhood where the threshold is met or exceeded. These fractions, sometimes termed neighborhood probabilities, are then used to compare each point in the forecast grid to its corresponding observation using the FSS.

Model forecasts are verified for four precipitation thresholds (0.254, 2.54, 6.35, and 12.7 mm h⁻¹) and six spatial scales (square neighborhood widths of 1, 3, 11, 21, 31, and 41 grid points). The precipitation thresholds range from light to heavy hourly precipitation rates. The lightest threshold of 0.254 mm h⁻¹ will capture broader regions of precipitation that characterize the rainfall footprint of precipitation systems. Higher thresholds will focus on smaller local-scale regions of moderate to heavy rainfall embedded within convective thunderstorms and precipitation systems. To aid in understanding of the six spatial scales, the neighborhood widths are presented in physical units based on the resolution of the verification grid (~4.763 km). Specifically, the approximate width of half the neighborhood square is used in place of the number of grid points, as shown in Table 5.1.

Table 5.1: Naming convention for physical neighborhood widths instead of grid points.

Neighborhood width (grid points)	Physical width convention (km)
1	0*
3	7
11	25
21	50
31	75
41	100

*0 km is used since 1 grid point is equivalent to no neighborhood being used

NWP model forecasts are verified over a specified domain for each of the three installations. Each domain is drawn so that forecast verification incorporates as much NWP model data as possible (including downscaled forecasts from the WRF simulations). The domains are kept large enough to characterize the overall precipitation forecast performance for each climatic region (i.e., Southeast, Mountain West, and Midwest) and to ensure verification scores are not unduly influenced by small sample size (i.e., a few particularly good or bad forecasts). Figure 5.2a shows the verification domain for Fort Hood. The zoomed view in Figure 5.2b shows the Fort Hood watershed boundary with the NCEP Stage IV verification grid overlaid. Also shown are examples of select neighborhood widths (0, 25, and 75 km) used to compute FSSs. Note that the neighborhoods are not fixed at these positions as verification occurs by moving across the entire domain. The fixed positions are shown only to depict the spatial scales relative to each watershed. The smallest width of 1 grid point (red) is used to compute the grid-scale accuracy of the precipitation forecasts. The 25-km neighborhood (green) is more representative of precipitation placement at the scale of the watershed itself. The 75-km neighborhood (blue) is used to verify forecasts at a scale slightly larger than the watershed. This analysis conveys the degree of accuracy in placing precipitation in the vicinity of the watershed but not necessarily inside it. In a similar fashion, Figure 5.3 and Figure 5.4 show the verification domains for FLW and Fort Carson, respectively.

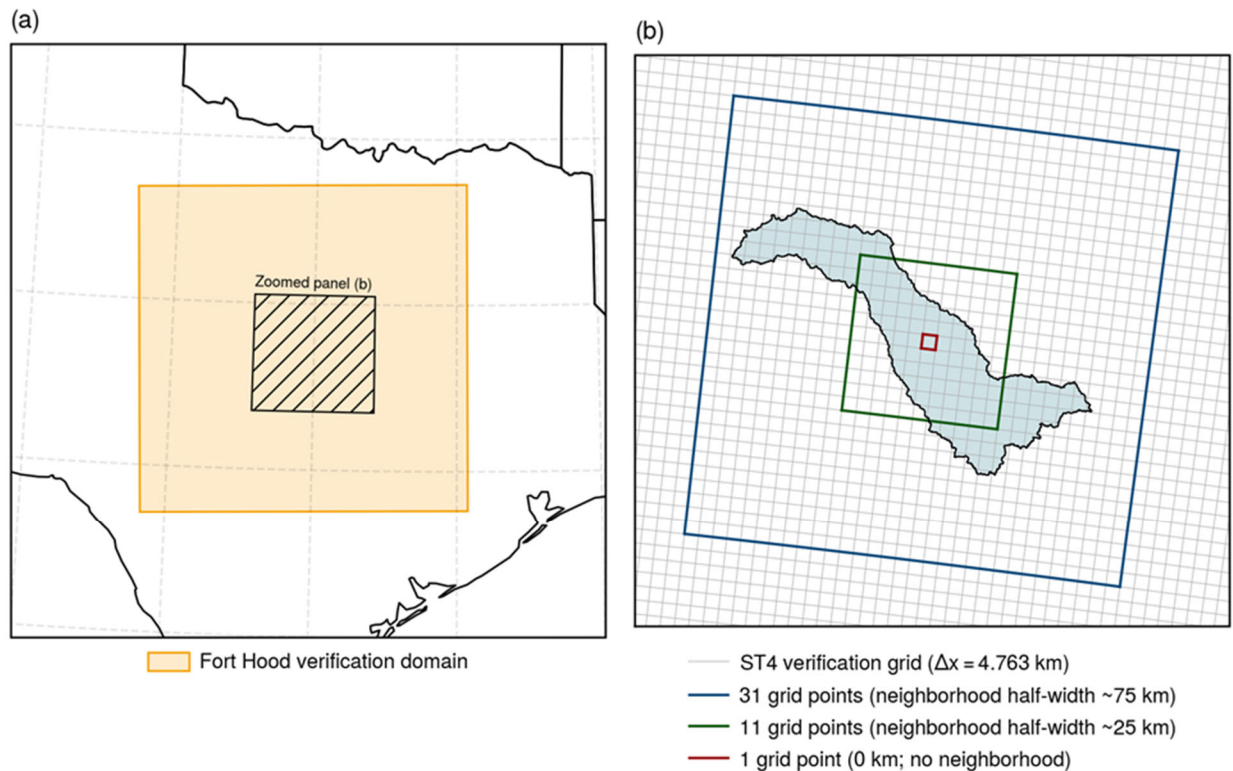


Figure 5.2: The Fort Hood domain used for precipitation forecast verification.

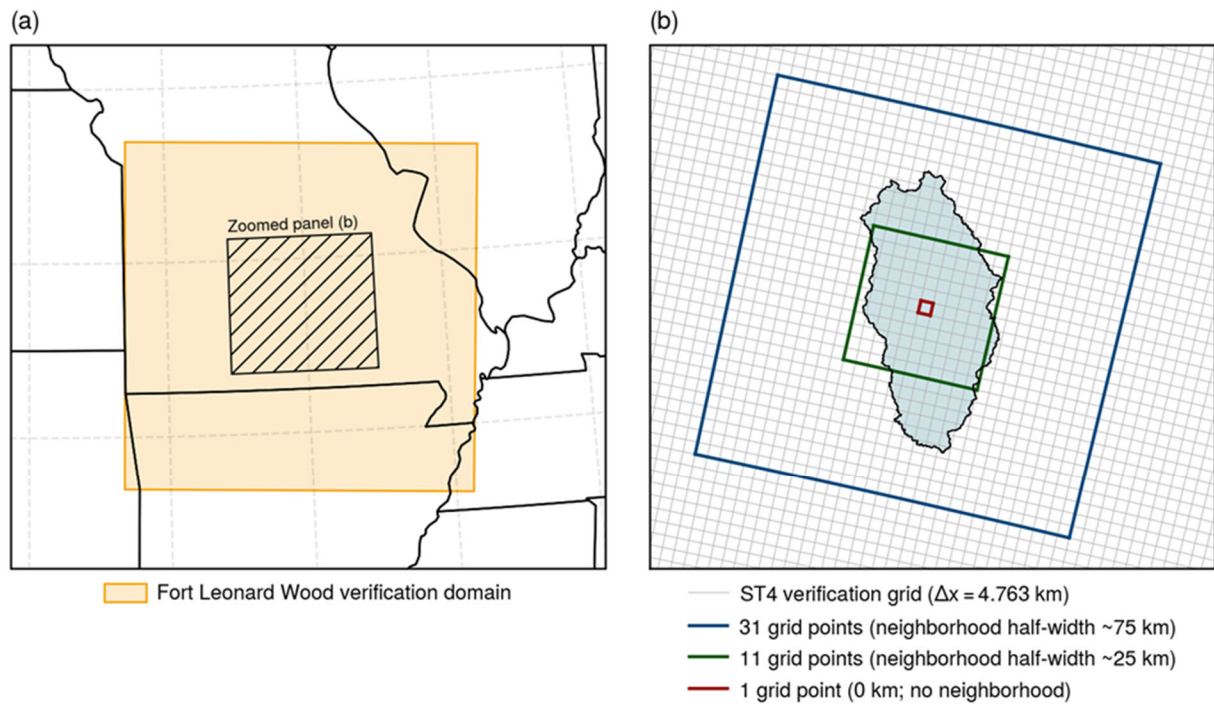


Figure 5.3: The Fort Leonard Wood domain used for precipitation forecast verification.

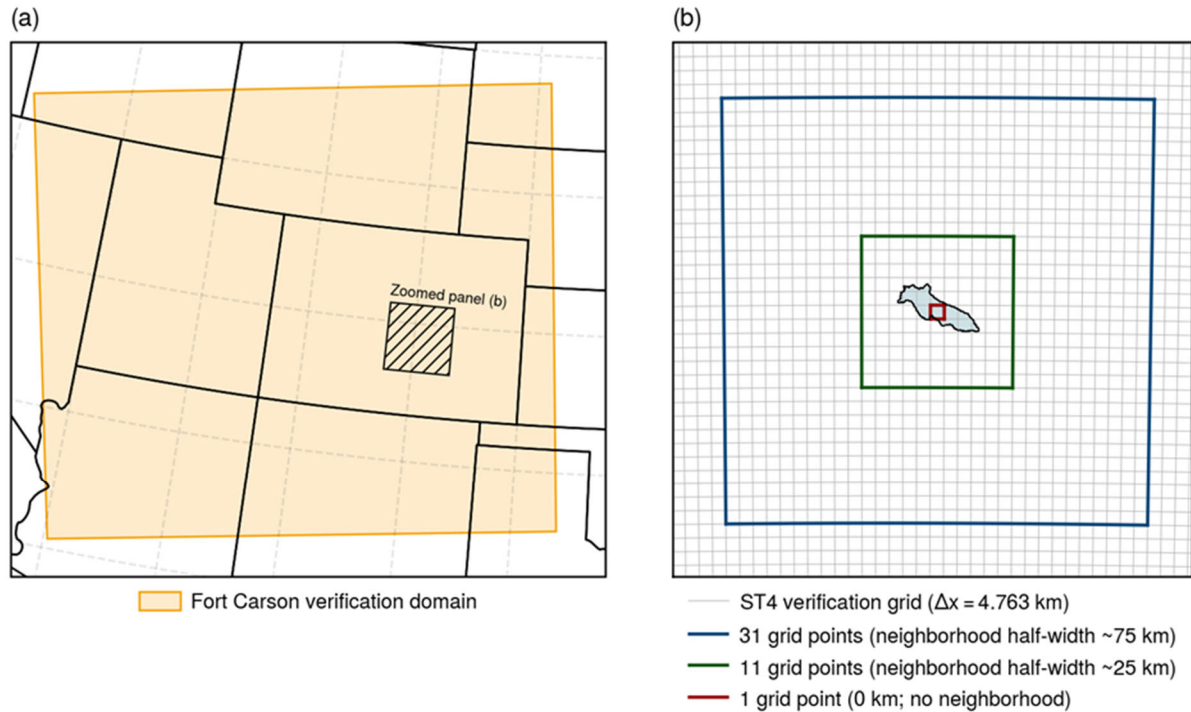


Figure 5.4: The Fort Carson domain used for precipitation forecast verification.

5.3.3 Hydrology Models

HEC-HMS accomplishes the physical representation of a watershed with a basin-scale model which simulates hydrologic processes such as soil infiltration, overland flow, and hydrologic routing. Widely available GIS data (i.e. topographic, soils, land use, and drainage networks) are used to create a physically based representation of a basin. HEC-HMS accepts gridded, long-term precipitation data which is transformed into saturation and infiltration of excess runoff at the gridded sub-basin scale using the gridded Soil Moisture Accounting (SMA) method. The excess runoff is routed overland through the ModClark transform method and then into the channel reaches to the basin outlet using the Muskingum-Cunge method.

The catchment and stream network for the models were delineated from the USGS NED (Gesch et al. 2002) digital elevation model (DEM). The minimum contributing area to initiate a stream was set to the default for each study subbasin (1% of the largest contributing area within each subbasin). For example, the Cowhouse-Owl HMS model at Fort Hood contains four subbasins (Cowhouse, Table Rock, House, and Owl) and streams were initiated within each basin depending on the size of the largest contributing area within each basin.

The necessary initial parameters for the SMA method (Figure 5.5) were estimated following Ahbari et al. (2018), Bennett and Peters (2000), and Holberg (2014). Saturated water content (field capacity), residual water content (wilting point), saturated hydraulic conductivity, and unsaturated hydraulic conductivity grids were generated using the NRCS Soil Survey Geographic database (SSURGO; NRCS 2019) and the U.S. Department of Agriculture (USDA) Agricultural Research Service (ARS) ROSETTA hydraulic parameter model (Schaap et al. 2001). Sand/silt/clay percentage, bulk density at 15 bars, and water content percentage at 15 and 33 bars were extracted from the SSURGO database and imported into ROSETTA with saturated water content, residual water content, saturated hydraulic conductivity, and unsaturated hydraulic conductivity as ROSETTA output. Soil storage and tension storage were estimated as the products of porosity and soil depth and of field capacity and soil depth, respectively. Soil depth was estimated as the lesser of root zone depth or soil restrictive layer depth, where root zone depth was estimated from land cover type (NRCS, 1997) and soil restrictive layer depth was determined from the SSURGO database. Groundwater 1 and 2 percolation values are determined during calibration, but initial values were set at 50% and 10% of unsaturated hydraulic conductivity, respectively.

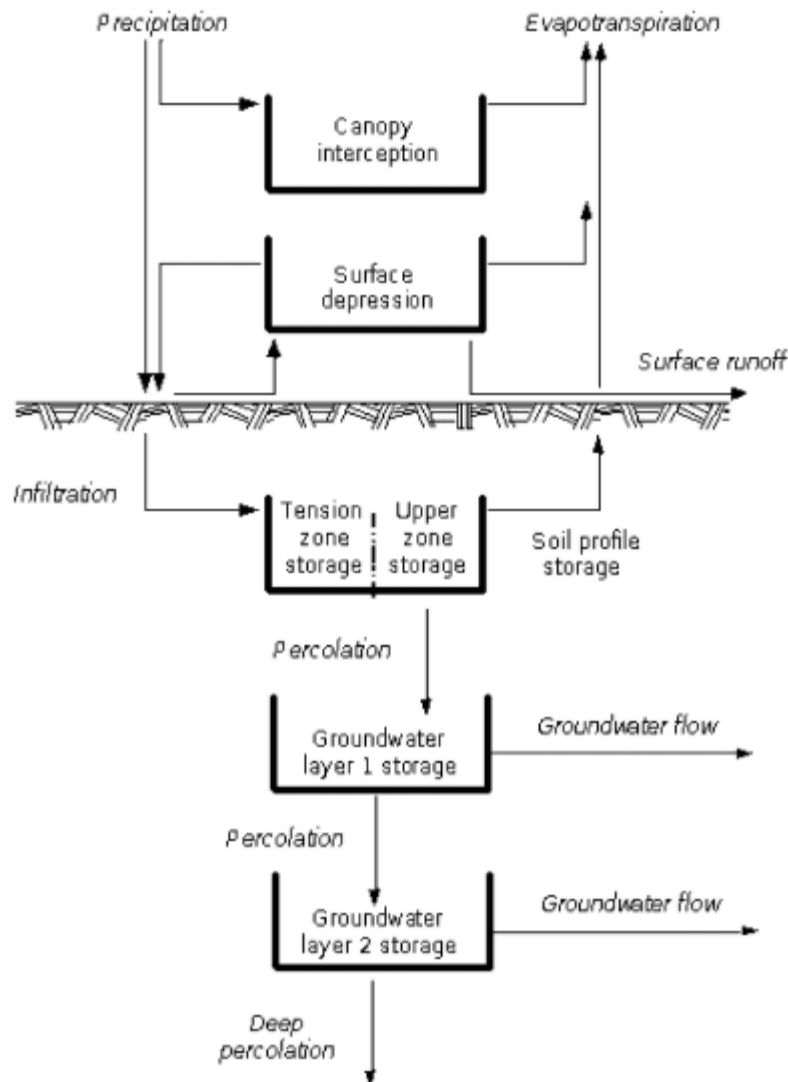


Figure 5.5: Conceptual schematic of the continuous Soil Moisture Accounting method (USACE, 2000)

The four additional groundwater parameters, GW1 storage, GW1 coefficient, GW2 storage, and GW2 coefficient were estimated graphically by flow recession analysis developed by Linsley (1958). Hydrographs were plotted using daily USGS gauge data from each basin for several isolated flow events to determine the separation of runoff, interflow, and baseflow (Figure 5.6). GW 1 parameters represent interflow and GW 2 parameters represent baseflow (groundwater).

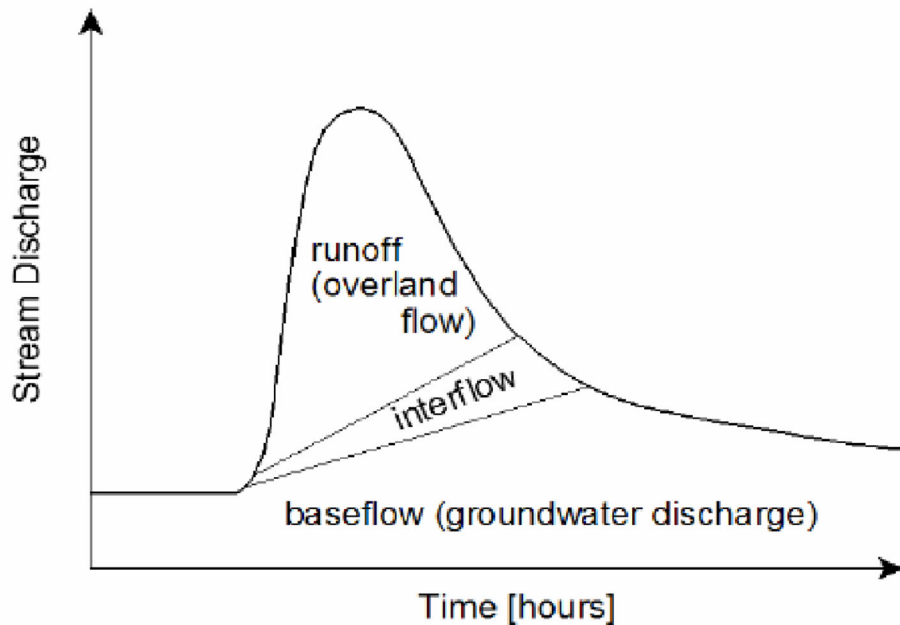


Figure 5.6: Conceptual schematic of the hydrologic response function of a basin, reflecting the separation of runoff, interflow, and baseflow (Harter et al., 2004).

Surface depression storage was determined based on surface slope calculated from the NED DEM and recommendations from Bennett (1998) (Table 5.2). Canopy storage was calculated from NLCD land cover data and recommendations from (USACE, 2000) (Table 5.3).

Table 5.2: Surface depression storage parameter values based on surface slope (Bennett, 1998).

Description	Slope (%)	Surface Storage (mm)
Paved Impervious Areas	NA	3.18-6.35
Flat, Furrowed Land	0-5	50.8
Moderate to Gentle Slope	5-30	6.35-12.70
Steep, Smooth Slopes	>30	1.02

Table 5.3: Vegetation canopy storage parameter values based on land cover type (USACE, 2000).

Land Cover	Canopy Storage (mm)
Open Water	0
Developed	0
Barren	0
Deciduous Forest	2
Evergreen Forest	2.5
Mixed Forest	2.2
Shrub/Scrub	1.27
Herbaceous	1.27
Pasture/Hay	1.27
Cultivated Crops	1.27
Woody Wetlands	2.2
Herbaceous Wetlands	1.27

HEC-HMS requires the ModClark transform method be used with a gridded Soil Moisture Accounting method. The ModClark method has two parameters, time of concentration and storage coefficient, that were estimated using the SCS watershed lag method (Mockus 1961; NRCS 2010). The Muskingum–Cunge routing method was selected for its applicability in basins that are difficult to calibrate due to the lack of hydrograph observations. Routing parameterization was completed based on the NED DEM and data obtained during site visits.

HEC-HMS models are calibrated using gauge precipitation data or gridded precipitation data (e.g. radar data) in conjunction with stream gauge data to validate simulation hydrographs with observed hydrographs for historic events. To calibrate and assess the continuous hydrologic models they need to be forced with precipitation and evapotranspiration (ET) data that is as close to reality as possible. Otherwise uncertainty in the precipitation and ET forcing would influence the comparison of streamflow observations and simulations, especially over long simulation time frames. Therefore, the HEC-HMS models were forced with the NCEP Stage IV (Du 2011) quantitative precipitation estimation (QPE) and gridded ET estimates from the mosaic North American Land Data Assimilation System (NLDAS-2: Xia et al. 2012). The Stage IV precipitation analysis is available on a 4-km grid across CONUS and at 1 hr temporal resolution. The NLDAS ET is available on a 1/8th degree grid across CONUS and at 1 hr temporal resolution. HEC-HMS streamflow simulations, with Stage IV precipitation and NLDAS ET forcing, were used to calibrate the models and then compared to USGS streamflow observations to assess hydrologic model performance.

The initial soil water content (initial soil storage) and initial GW1 storage, GW2 storage, canopy storage, and surface storage parameter values (as a percent of total depth) were determined using a warmup period on the HMS models. Each model (Hood, FLW, and Carson) were run for a 2-yr simulation, ending 31 Dec 2015, to determine storage parameter values for that date. The models were again run starting the first of each month in 2015 with various initial storage values (%) to determine the optimum initial values and required warmup simulation length needed for the

storage values to converge with the original long-term simulation. This analysis determined that soil storage required the most simulation time to converge for all models, with Hood, FLW, and Carson requiring warm up periods of seven, six, and eight months, respectively. An eight-month warmup period was used for all calibration simulation runs for all models.

HEC-HMS includes model optimization algorithms to calibrate models so that simulated results match observed streamflow as closely as possible. The optimization algorithm works by adjusting parameter values and initial conditions, including Manning's n for the routing method and the ModClark parameter values. Unfortunately, when the gridded SMA method is used, the optimization algorithms can only adjust the five storage initial conditions and the non-gridded GW parameters, but not of the gridded parameter values (cover storage, surface storage, GW1 and GW2 storage, GW1 and GW2 coefficients, soil and tension storage, and saturated and unsaturated hydraulic conductivity).

In order to calibrate HEC-HMS models utilizing the gridded SMA method a script was developed to run the HEC-HMS models while changing one or more parameter values each run and saving the output. A range of gridded SMA parameter files were created for each parameter, ranging from 5% of the original estimated value(s) to 500%. Since tension storage must be smaller than soil storage, tension storage files with values ranging from 5% to 100% of the original estimated values and soil storage files with values ranging from 100% to 500% of the original estimated values were created. The HEC-HMS models were run while changing one parameter at a time for the full range of values to determine the model sensitivity to each parameter and the outputs were compared to the available observed data both graphically and statistically. Many of the model parameters are physically based and a wide range of possible values were outside of feasible parameter limits, however, the entire range of parameter values were included to better assess model sensitivity to all parameters.

The initial parameter sensitivity assessment for all three models indicated that the ratio between soil storage and tension storage greatly affected hydrograph peaks and attenuation. The three GW1 parameters and three GW2 parameters affect the shape of the falling limb of the hydrographs and the quantity and persistence of baseflow, respectively. The ModClark parameters (time of concentration and storage coefficient) affect the timing, shape, and peaks of the hydrographs. Calibration was limited to these parameters because most (other than tension storage) are classically used as calibration parameters while most other parameters are physically based.

Hydrologic model performance was assessed in three regards following the outline by Ritter and Muñoz-Carpena (2013). The assessments include a graphical representation of the relationship between model simulations and observations, an absolute value error indicator, and a dimensionless index for quantifying the goodness of fit. We used the root-mean-square error (RMSE), RMSE-observations standard deviation ratio (RSR), NSE coefficient, and KGE coefficient. Calibration was conducted by first adjusting tension storage and evaluated graphically and statistically. Once a tension storage correction was set, the six GW parameters were used to adjust the falling limb of the flood hydrographs and the quantity and persistence of baseflow and compared graphically to the observed streamflow. Finally, all combinations of the ModClark parameters were tested to determine the statistically optimum combination across the parameter space (Figure 5.7).

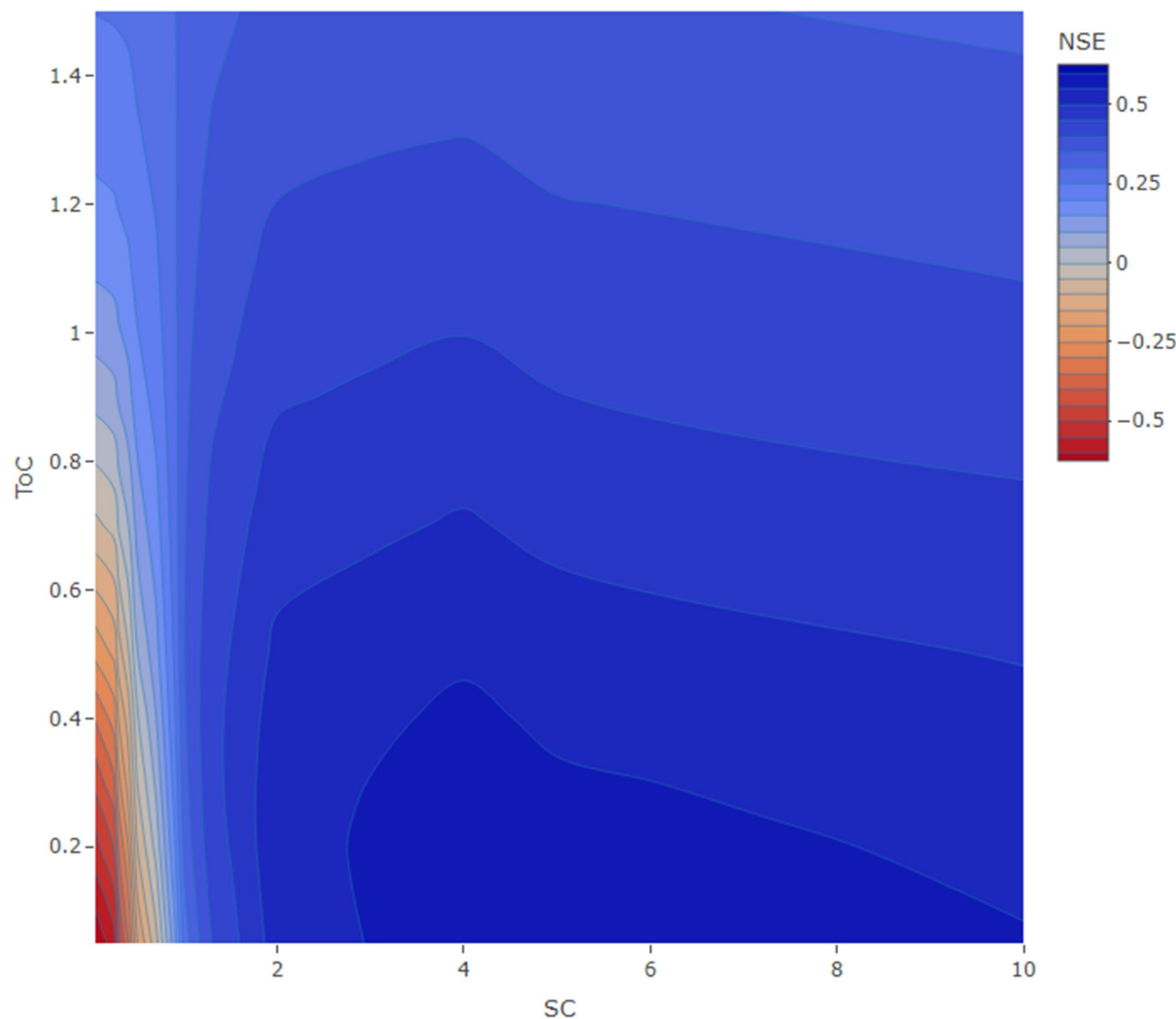


Figure 5.7: Example heatmap of NSE across ModClark parameter space for Fort Hood calibration period.

Calibration and validation simulations are two calendar years long. The calibration and validation periods for Hood and FLW are 2016-2017 and 2018-2019, respectively, with 2015 as the warmup period. At Fort Carson 2019 was a very dry year and no observed streamflow data was recorded. As a result, the calibration and validation periods for Carson are 2015-2016 and 2017-2018, respectively, with 2014 as the warmup period.

Visually, the Fort Hood calibrated streamflow simulation matches the observations well with regards to rising and falling limb timing (Figure 5.8). Simulated peak flows are mixed compared to the observations, with some events being over or underestimated, though most of the underestimated events are those with observed peak flow greater than 200 cu. m per sec. Generally, the calibration process reduced simulated peak flow values compared to the uncalibrated simulations. The Validation period shows similar results with regards to simulated flow event timing compared to observations (Figure 5.9). The greatest discrepancy in timing is during a March 2019 event (Figure 5.9b), when the simulation rising limb occurs several hours before the

observed. Simulated peak flows are again mixed compared to the observations but with the simulated peak flow of the largest event greater than the observed (Figure 5.9a) and underestimated during small events (Figure 5.9d).

Statistically the calibrated Fort Hood HMS model is acceptable across both the absolute value error statistics and the normalized goodness-of-fit statistics (Table 5.4). The calibration process improved all statistics when considering the full calibration period and the individual calendar years, except 2016 and 2019, which showed decreases in efficiency. Nonetheless the KGE values indicate that the model improves upon the mean flow benchmark.

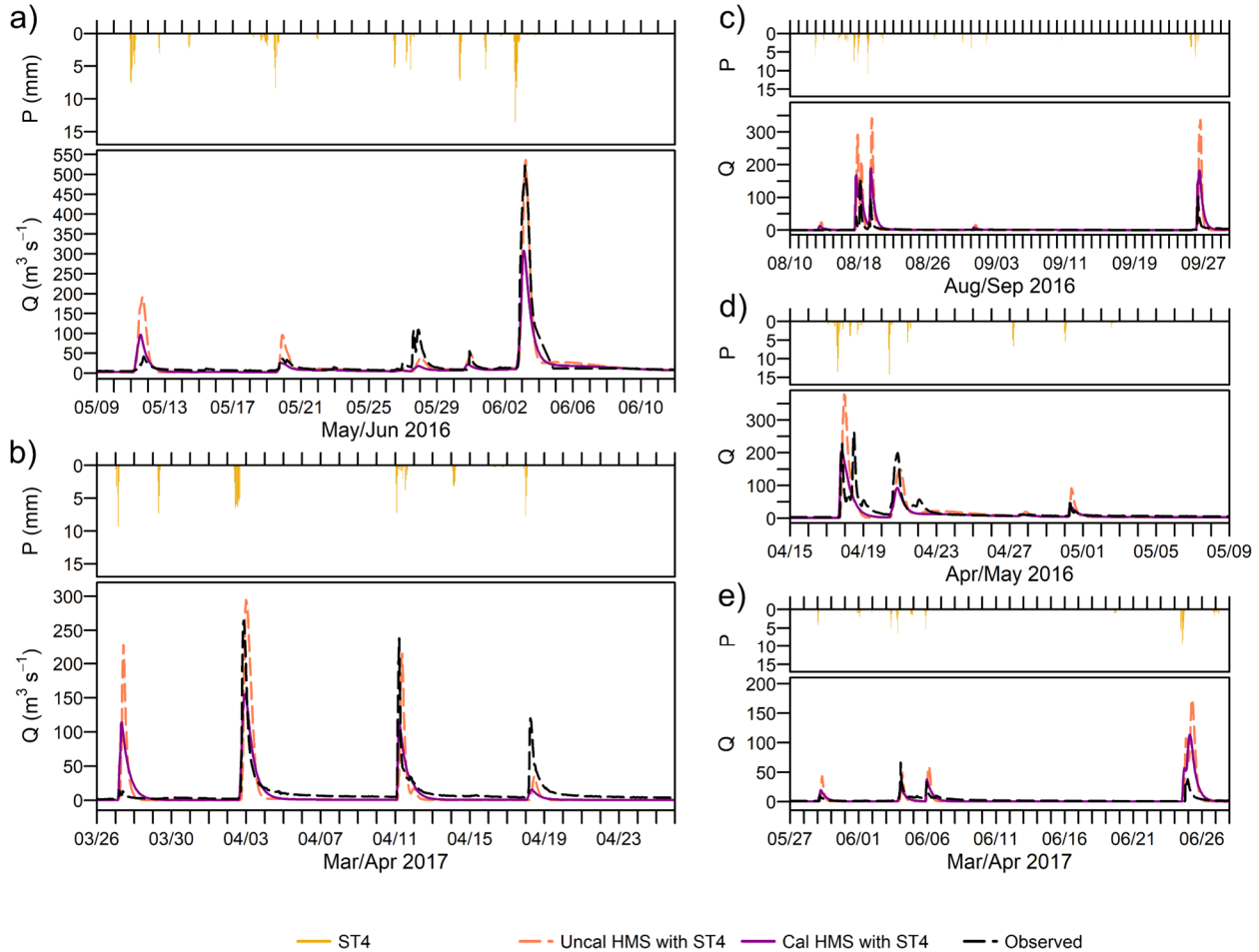


Figure 5.8: Graphs of the uncalibrated and calibrated HEC-HMS performance at the Fort Hood USGS gauge during various timeframes of the 2016-2017 calibration period. (a) The top plot is the hyetograph of hourly Stage IV precipitation averaged across the area of the Cowhouse Creek catchment upstream of the Pidcoke USGS stream gauge. The bottom plot shows the observed streamflow (black dashed line) from the USGS gauge and the uncalibrated (orange dashed line) and calibrated (purple line) HEC-HMS modeled streamflow from long term simulations forced by the Stage IV precipitation forced in the top plot. (b)-(e) as in (a), but for other timeframes.

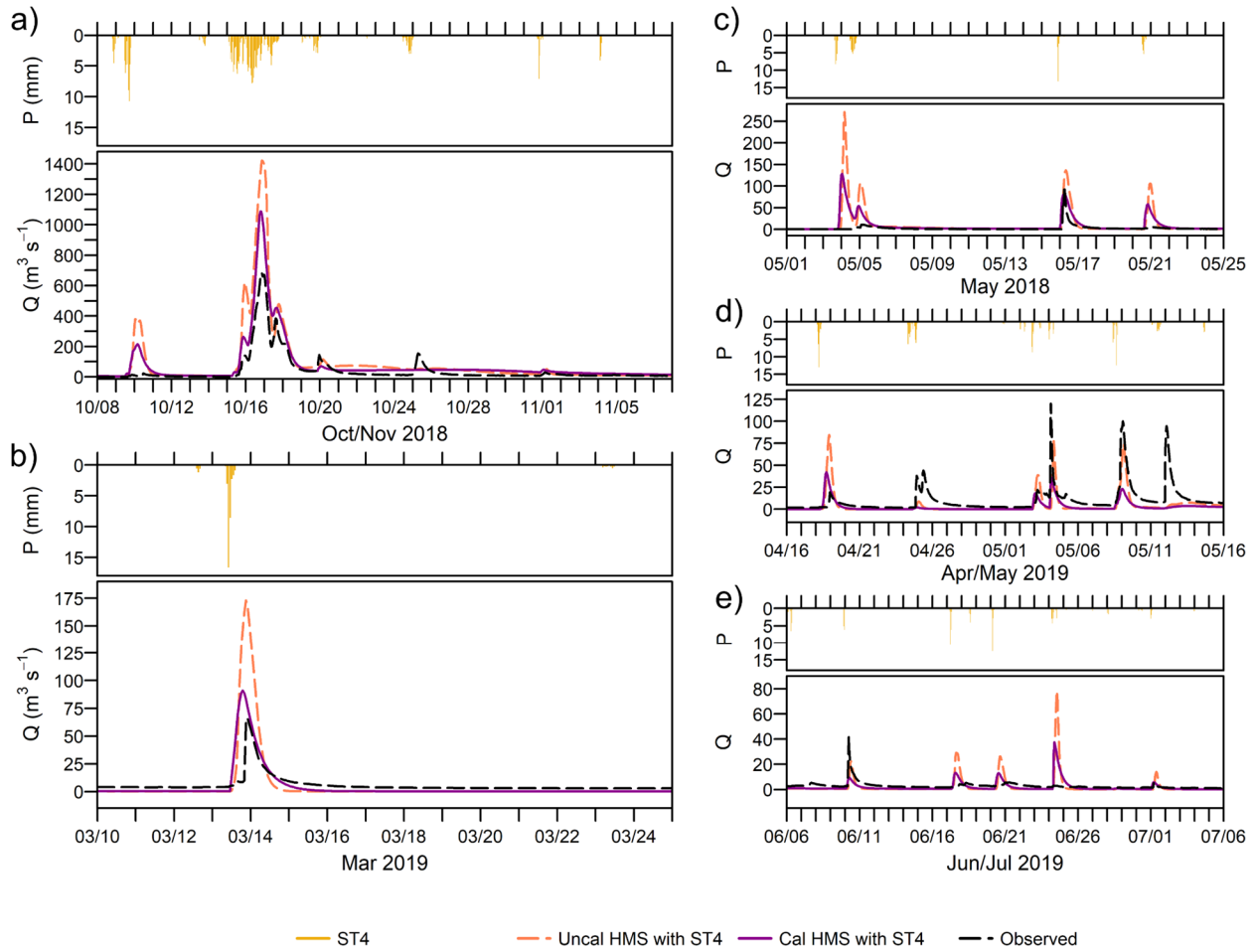


Figure 5.9: Graphs of the uncalibrated and calibrated HEC-HMS performance at the Fort Hood USGS gauge during various timeframes of the 2018-2019 validation period. (a) The top plot is the hyetograph of hourly Stage IV precipitation averaged across the area of the Cowhouse Creek catchment upstream of the Pidcock USGS stream gauge. The bottom plot shows the observed streamflow (black dashed line) from the USGS gauge and the uncalibrated (orange dashed line) and calibrated (purple line) HEC-HMS modeled streamflow from long term simulations forced by the Stage IV precipitation forced in the top plot. (b)-(e) as in (a), but for other timeframes.

Table 5.4: Model evaluation statistics for the uncalibrated and calibrated Cowhouse Creek HEC-HMS performance at the Fort Hood Pidcoke USGS gauge during the 2016-2017 calibration period and 2018-2019 validation period and each calendar year.

		Calibration Period			Validation Period		
		2016-2017	2016	2017	2018-2019	2018	2019
NSE	Uncal	0.21	0.30	-0.43	-0.59	-0.64	0.13
	Calibrated	0.58	0.60	0.43	0.47	0.49	0.17
KGE	Uncal	0.60	0.65	0.37	-0.27	-0.80	0.36
	Calibrated	0.62	0.62	0.60	0.35	-0.14	0.09
RSR	Uncal	0.89	0.84	1.20	1.26	1.28	0.93
	Calibrated	0.65	0.63	0.75	0.73	0.72	0.91
RMSE	Uncal	14.91	18.18	10.68	29.84	41.53	7.46
	Calibrated	10.80	13.70	6.74	17.24	23.26	7.30

The FLW HMS model included both the Big Piney River and Roubidoux Creek, but each basin was calibrated separately using two different USGS stream gauges. The calibration results for each are discussed apart. Visually, the calibrated Big Piney River streamflow simulation matches the observations well with regards to rising, falling limb timing, and baseflow (Figure 5.10). Simulated peak flows are mixed compared to the observations, with some events being underestimated, though most of the underestimated events are those with observed peak flow less than 150-200 cu. m per sec. Generally, the calibration process reduced simulated peak flow values compared to the uncalibrated simulations but also greatly improved the falling limb and baseflow compared to the observations. The Validation period shows similar results with regards to simulated flow event timing, falling limb timing, and baseflow drawdown compared to observations (Figure 5.11). Simulation peak flows closely match the observations during some large events (Figure 5.11b) but underestimate the February 2018 event (Figure 5.11a). The February 2018 discrepancy in peak flow is likely due to winter conditions such as rain on snow or frozen ground which are not fully represented in the HMS model.

Statistically the calibrated Big Piney HMS basin model is acceptable across both the absolute value error statistics and the normalized goodness-of-fit statistics (Table 5.5). The calibration process improved all statistics when considering the full calibration period and the individual calendar years. The calendar year with the lowest combined statistical

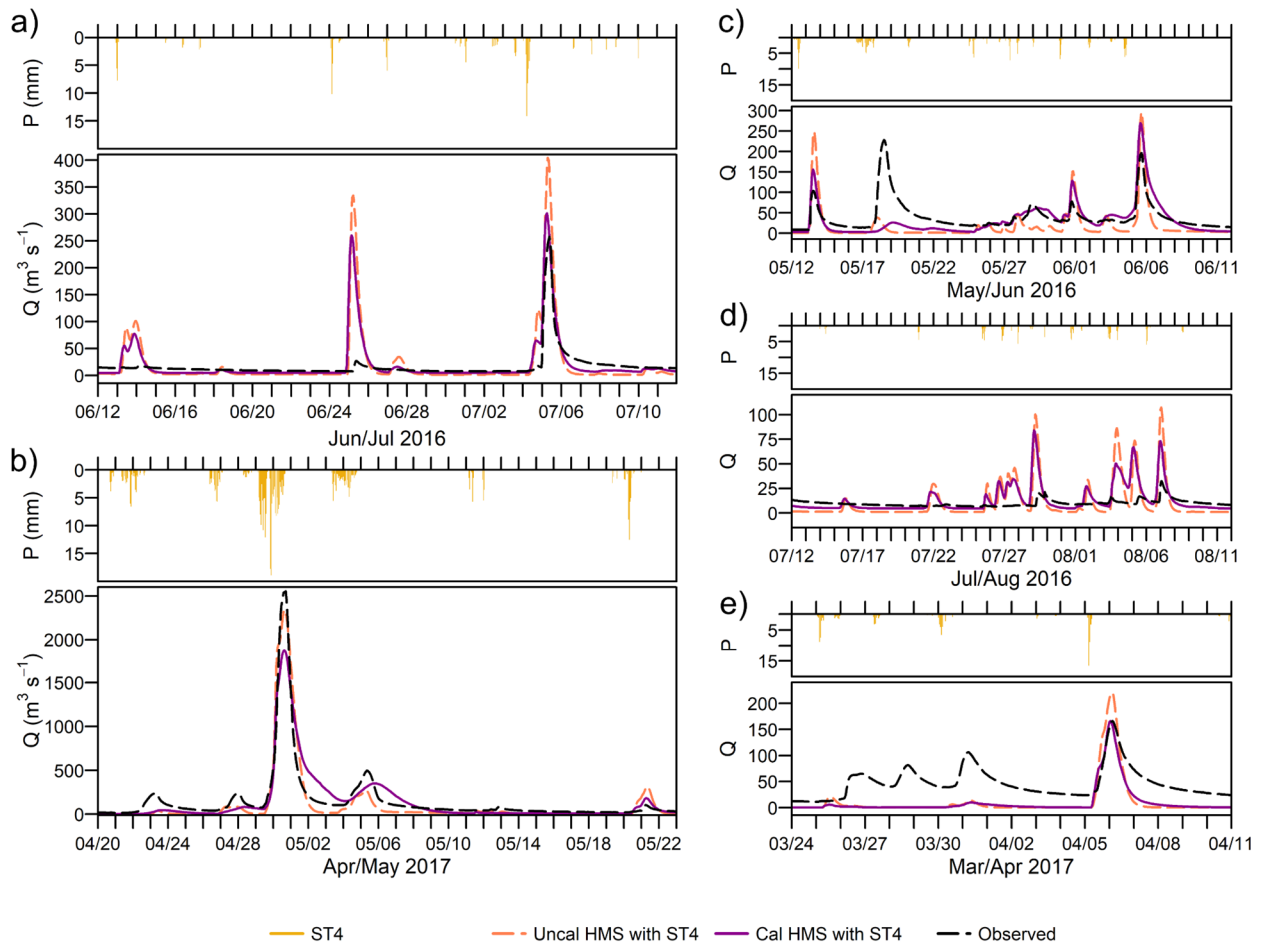


Figure 5.10: Graphs of the uncalibrated and calibrated HEC-HMS performance at the FLW USGS gauge during various timeframes of the 2016-2017 calibration period. (a) The top plot is the hyetograph of hourly Stage IV precipitation averaged across the area of the Big Piney River catchment upstream of the USGS stream gauge near Big Piney, MO. The bottom plot shows the observed streamflow (black dashed line) from the USGS gauge and the uncalibrated (orange dashed line) and calibrated (purple line) HEC-HMS modeled streamflow from long term simulations forced by the Stage IV precipitation forced in the top plot. (b)-(e) as in (a), but for other timeframes.

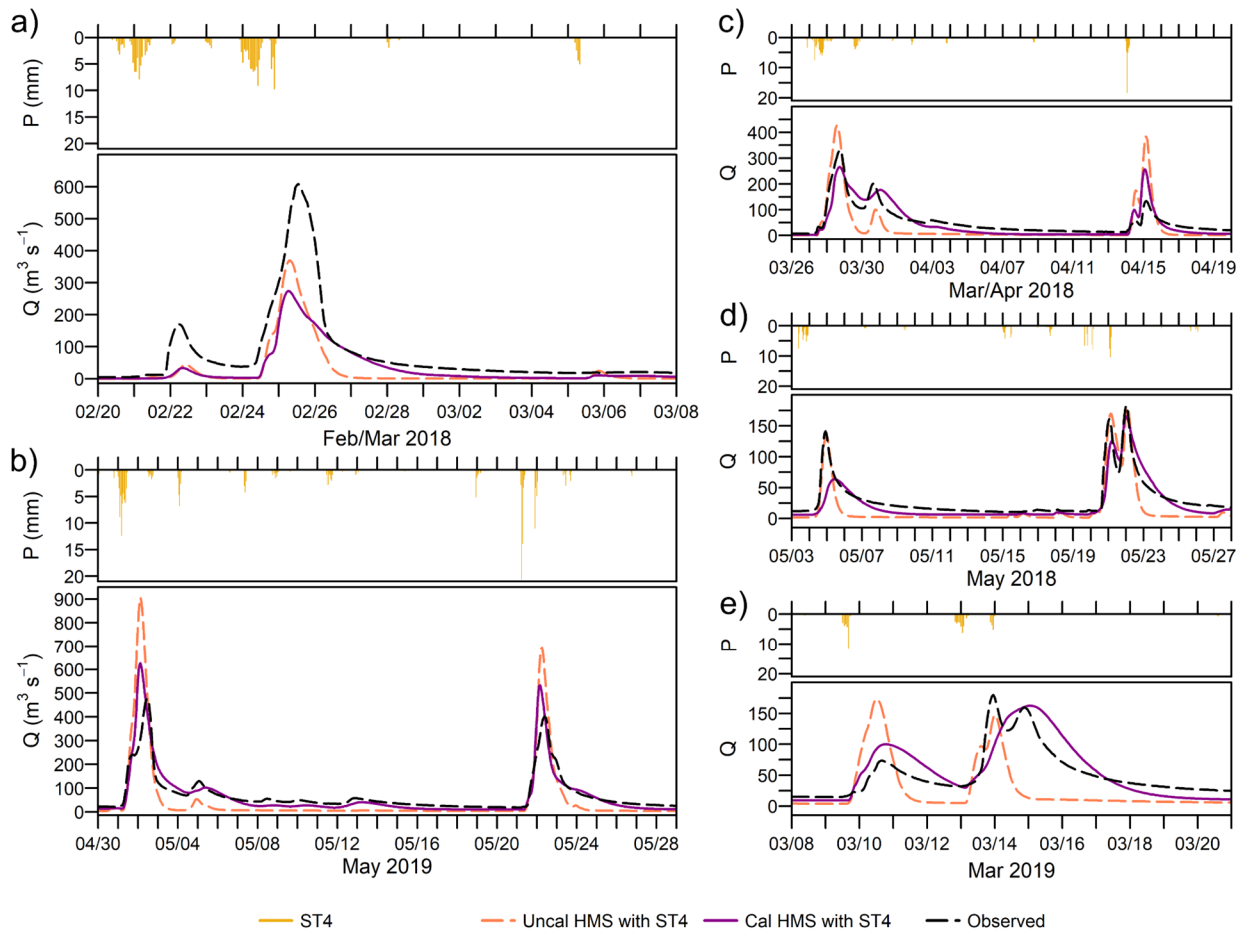


Figure 5.11: Graphs of the uncalibrated and calibrated HEC-HMS performance at the FLW USGS gauge during various timeframes of the 2018-2019 validation period. (a) The top plot is the hyetograph of hourly Stage IV precipitation averaged across the area of the Big Piney River catchment upstream of the USGS stream gauge near Big Piney, MO. The bottom plot shows the observed streamflow (black dashed line) from the USGS gauge and the uncalibrated (orange dashed line) and calibrated (purple line) HEC-HMS modeled streamflow from long term simulations forced by the Stage IV precipitation forced in the top plot. (b)-(e) as in (a), but for other timeframes.

Table 5.5: Model evaluation statistics for the uncalibrated and calibrated Big Piney River HEC-HMS performance at the FLW USGS stream gauge near Big Piney, MO during the 2016-2017 calibration period and 2018-2019 validation period and each calendar year.

		Calibration Period			Validation Period		
		2016-2017	2016	2017	2018-2019	2018	2019
NSE	Uncal	0.86	-0.72	0.90	-0.08	0.24	-0.50
	Calibrated	0.86	0.08	0.88	0.47	0.52	0.38
KGE	Uncal	0.53	0.17	0.57	0.31	0.38	0.16
	Calibrated	0.72	0.52	0.71	0.71	0.61	0.59
RSR	Uncal	0.37	1.31	0.31	1.04	0.87	1.23
	Calibrated	0.38	0.96	0.35	0.73	0.69	0.79
RMSE	Uncal	29.86	23.29	35.24	37.12	32.54	41.20
	Calibrated	30.43	17.04	39.54	26.11	25.81	26.41

Visually, the calibrated Roubidoux Creek streamflow simulation matches the observations well with regards to rising limb timing and baseflow, but the falling limb timing is often delayed (Figure 5.12). Baseflow and falling limb timing along Roubidoux creek varies from reach to reach due to the unlying karst topography. Simulated peak flows are mixed compared to smaller observed peaks of less than 100 cu. m per sec. The largest observed flood during the calibration period (Figure 5.12b) is also underestimated by the simulation. Generally, the calibration process reduced simulated peak flow values and falling limb drawdown timing compared to the uncalibrated simulations but also improved baseflow compared to the observations. The Validation period shows similar results with regards to simulated flow event timing and baseflow drawdown compared to observations but better falling limb timing compared to the calibration period (Figure 5.13). In general, the simulation during the validation period underestimates peak flow for all observed events with peaks larger than 150 cu. m per sec. More large flow events occurred on Roubidoux Cr during the validation period than during the calibration period, which may partially explain the relatively poor performance of the calibrated model due to a limited number of large flow event data points. The greatest discrepancies in peak flow are during May 2019 events (Figure 5.13b), when the underestimates are not due to winter conditions. The repeated underestimated simulated peak flows are likely due to the calibration processes not adequately capturing and accounting for the basin's karst topography and resulting complex interactions between surface runoff, variable infiltration rates, near-surface groundwater.

Statistically the calibrated Roubidoux HMS basin model is acceptable across both the absolute value error statistics and the normalized goodness-of-fit statistics, other than the NSE for 2016, which was a dry year (Table 5.6). The calibration process treated both 2016 and 2017 with equal weight, resulting in an improvement in calibration statistics for 2016 but a slight worsening for 2017. In the future the number of, and relative size of, flow events during a prospective calibration period should be considered before selecting a given timeframe. More accurate simulated peak flows compared to observed may have been achieved by including 2015 in the calibration period. The calibration process improved all statistics when considering the whole validation period and the individual calendar years.

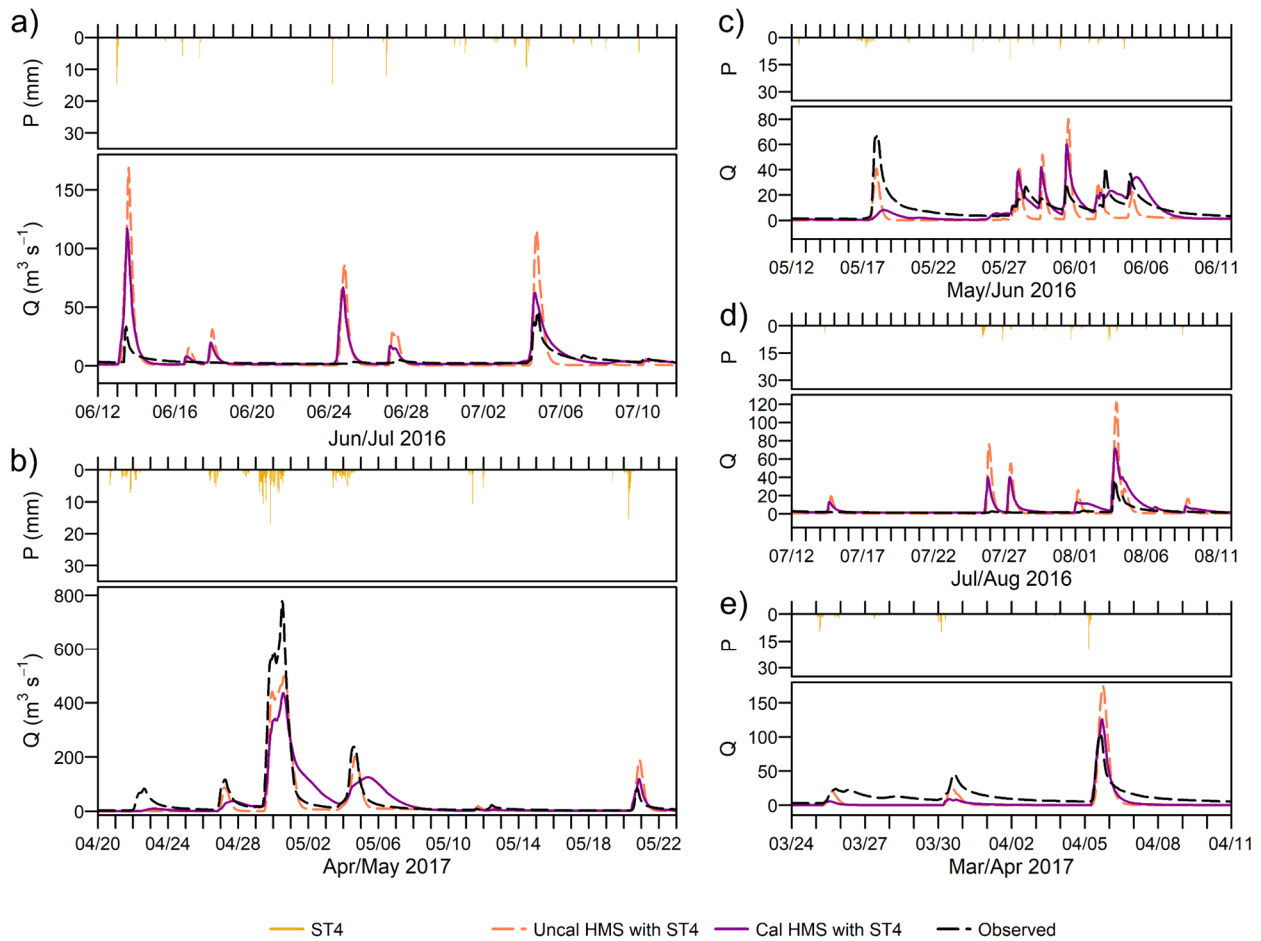


Figure 5.12: Graphs of the uncalibrated and calibrated HEC-HMS performance at the FLW USGS gauge during various timeframes of the 2016-2017 calibration period. (a) The top plot is the hyetograph of hourly Stage IV precipitation averaged across the area of the Roubidoux Creek catchment upstream of the USGS stream gauge above FLW, MO. The bottom plot shows the observed streamflow (black dashed line) from the USGS gauge and the uncalibrated (orange dashed line) and calibrated (purple line) HEC-HMS modeled streamflow from long term simulations forced by the Stage IV precipitation forced in the top plot. (b)-(e) as in (a), but for other timeframes.

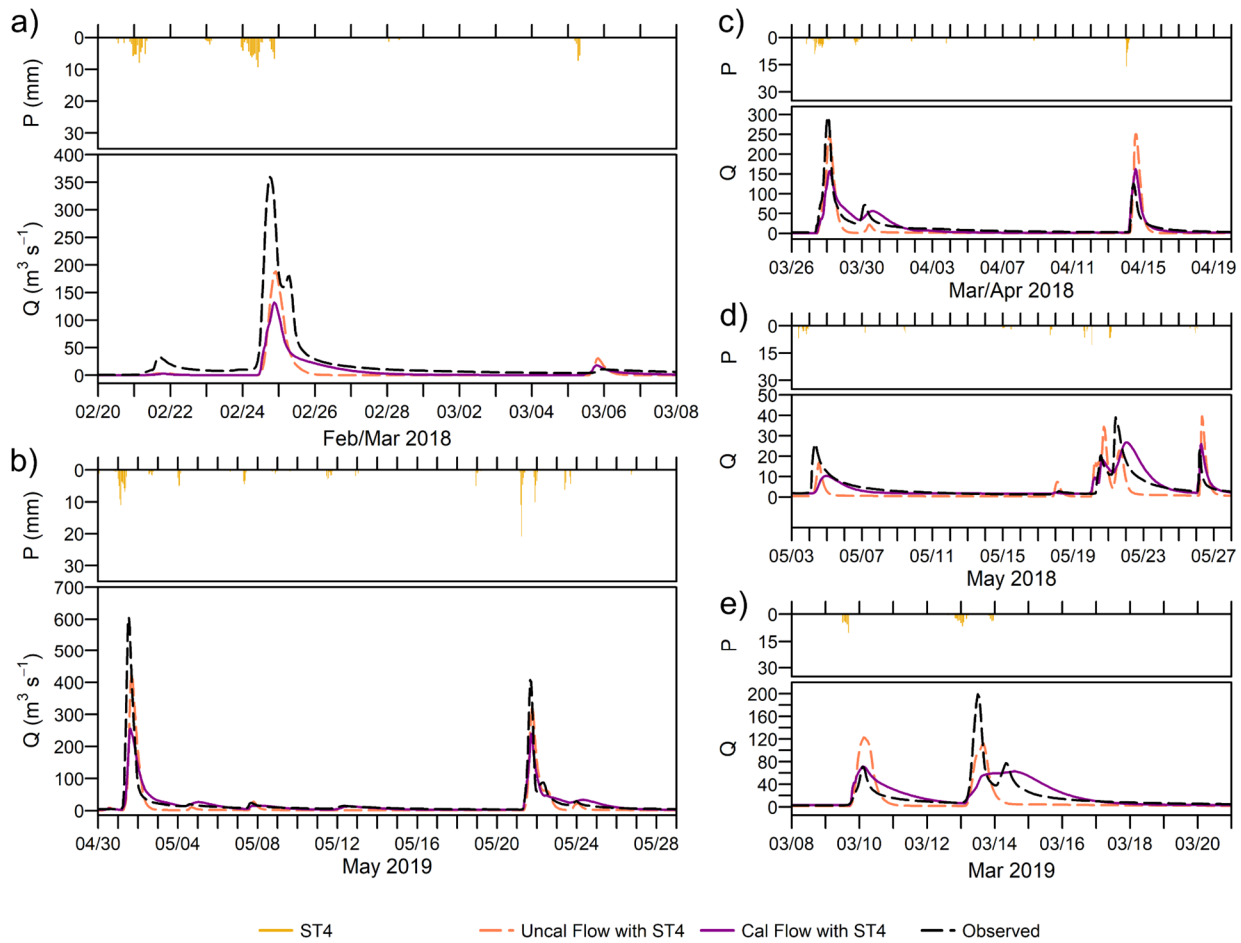


Figure 5.13: Graphs of the uncalibrated and calibrated HEC-HMS performance at the FLW USGS gauge during various timeframes of the 2018-2019 validation period. (a) The top plot is the hyetograph of hourly Stage IV precipitation averaged across the area of the Roubidoux Creek catchment upstream of the USGS stream gauge above FLW, MO. The bottom plot shows the observed streamflow (black dashed line) from the USGS gauge and the uncalibrated (orange dashed line) and calibrated (purple line) HEC-HMS modeled streamflow from long term simulations forced by the Stage IV precipitation forced in the top plot. (b)-(e) as in (a), but for other timeframes.

Table 5.6: Model evaluation statistics for the uncalibrated and calibrated Roubidoux Creek HEC-HMS performance at the FLW USGS stream gauge above FLW, MO during the 2016-2017 calibration period and 2018-2019 validation period and each calendar year.

		Calibration Period			Validation Period		
		2016-2017	2016	2017	2018-2019	2018	2019
NSE	Uncal	0.81	-2.13	0.86	0.18	0.00	0.29
	Calibrated	0.73	-0.52	0.75	0.41	0.33	0.46
KGE	Uncal	0.57	-0.16	0.53	0.45	0.42	0.46
	Calibrated	0.65	0.31	0.59	0.64	0.61	0.62
RSR	Uncal	0.43	1.77	0.38	0.91	1.00	0.84
	Calibrated	0.52	1.23	0.50	0.77	0.82	0.74
RMSE	Uncal	11.24	7.93	13.79	17.22	16.77	17.66
	Calibrated	13.46	5.54	18.22	14.62	13.75	15.44

Little Fountain Creek is an ephemeral stream with a high percentage of days with no flow conditions during the year. Keeton Reservoir, located just upstream of Fort Carson, is uncontrolled and attenuates flow downstream, especially following prolonged dry conditions when the reservoir is empty prior to the most recent rainfall event. Keeton Reservoir is included in the Little Fountain Creek HEC-HMS model. The stream experiences smaller, less numerous flow events than the other study basins in Missouri and Texas. The largest flow event 2015-2019 was 22 cu. m per sec. Due to the relatively small, infrequent observed flow events the graphical results of the model calibration were given more weight than the statistical analysis.

Visually, the calibrated Little Fountain Creek streamflow simulation matches the observations well with regards to rising limb timing and baseflow (Figure 5.14). Simulated peak flows and falling limb timing are mixed compared to the observations, with some simulated events being accurate (Figure 5.14a) and some underestimated (Figure 5.14b,c). Generally, the calibration process reduced simulated peak flow values compared to the uncalibrated simulations but also greatly improved the falling limb timing compared to the observations. The validation period shows the simulation peak flow underestimated during the only large event to occur 2017-2018 (Figure 5.15b).

Statistically the calibration process lowered the absolute value error statistics but the normalized goodness-of-fit statistics are mixed (Table 5.7). The highest performing year is 2015, which included the largest and more frequent flood events, indicating that the calibrated model could be relied upon to simulate larger flash flood events causing hazardous conditions in the basin. The calibration process generally improved the statistics when considering the full calibration period and the individual calendar years.

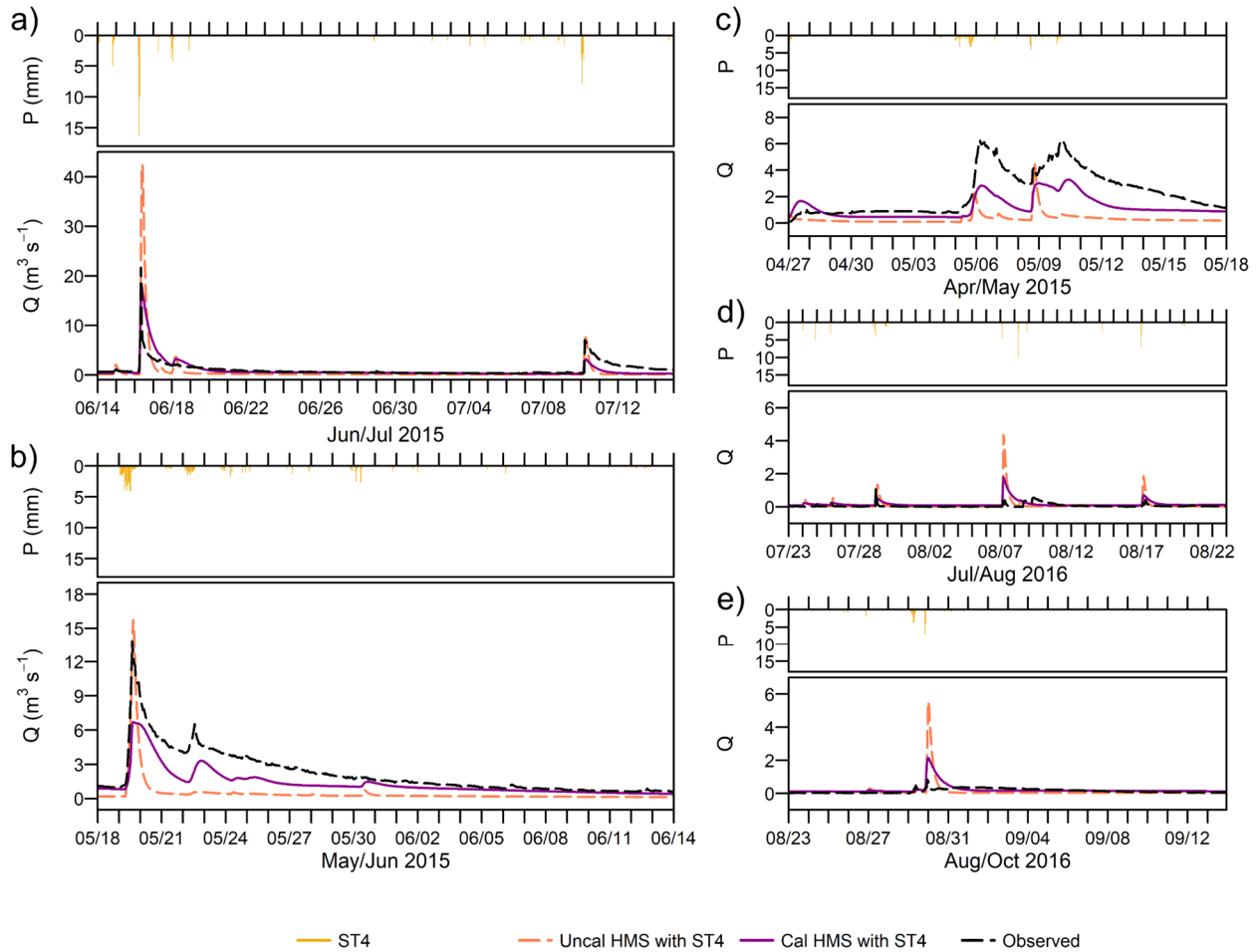


Figure 5.14: Graphs of the uncalibrated and calibrated HEC-HMS performance at the Fort Carson USGS gauge during various timeframes of the 2015-2016 calibration period. (a) The top plot is the hyetograph of hourly Stage IV precipitation averaged across the area of the Little Fountain Creek catchment upstream of the USGS stream gauge near Fountain, CO. The bottom plot shows the observed streamflow (black dashed line) from the USGS gauge and the uncalibrated (orange dashed line) and calibrated (purple line) HEC-HMS modeled streamflow from long term simulations forced by the Stage IV precipitation forced in the top plot. (b)-(e) as in (a), but for other timeframes.

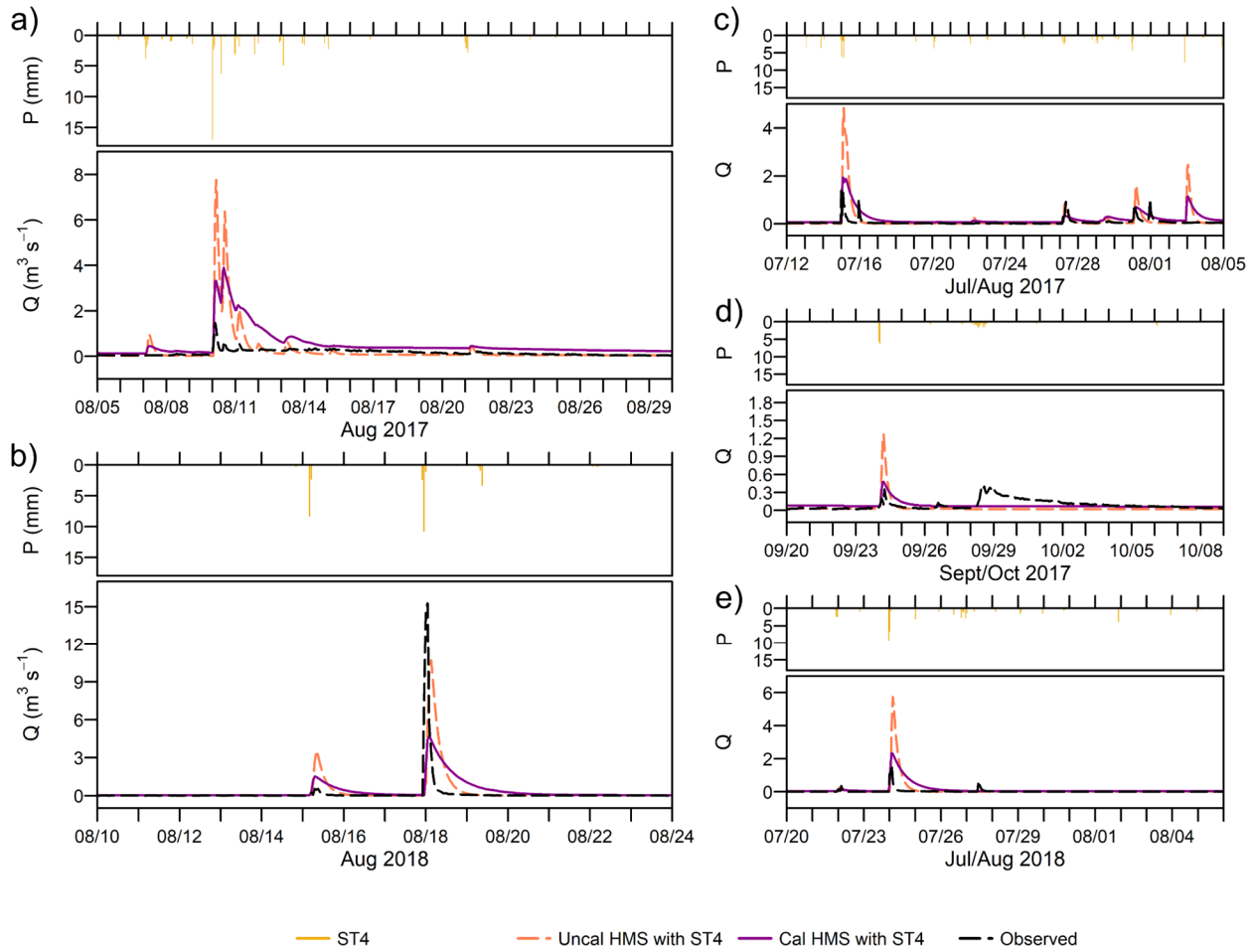


Figure 5.15: Graphs of the uncalibrated and calibrated HEC-HMS performance at the Fort Carson USGS gauge during various timeframes of the 2017-2018 validation period. (a) The top plot is the hyetograph of hourly Stage IV precipitation averaged across the area of the Little Fountain Creek catchment upstream of the USGS stream gauge near Fountain, CO. The bottom plot shows the observed streamflow (black dashed line) from the USGS gauge and the uncalibrated (orange dashed line) and calibrated (purple line) HEC-HMS modeled streamflow from long term simulations forced by the Stage IV precipitation forced in the top plot. (b)-(e) as in (a), but for other timeframes.

Table 5.7: Model evaluation statistics for the uncalibrated and calibrated Little Fountain Creek HEC-HMS performance at the Fort Carson USGS stream gauge near Fountain, CO during the 2015-2016 calibration period and 2017-2018 validation period and each calendar year.

		Calibration Period			Validation Period		
		2015-2016	2015	2016	2017-2018	2017	2018
NSE	Uncal	-0.23	-0.37	-4.19	-0.98	-9.09	-0.32
	Calibrated	0.62	0.59	-1.10	-0.27	-5.44	0.15
KGE	Uncal	0.16	0.12	-0.48	0.24	-1.54	-0.30
	Calibrated	0.62	0.56	-0.04	-0.20	-0.98	-1.34
RSR	Uncal	1.11	1.17	2.28	1.41	3.18	1.15
	Calibrated	0.61	0.64	1.45	1.13	2.54	0.92
RMSE	Uncal	1.11	1.55	0.25	0.35	0.31	0.38
	Calibrated	0.61	0.85	0.16	0.28	0.25	0.31

Hydro RF produces probabilistic streamflow and hazard forecasts by ingesting an ensemble of precipitation forecasts. The only variables which change from member to member are the timing and quantity of precipitation. Therefore, the sensitivity of the HEC-HMS models to changes in precipitation was assessed by measuring the percentage changes in the 7 largest individual peak flow events during 2020-2021 of the calibrated models due to percentage changes in the Stage IV precipitation (Figure 5.16). Results indicate that the calibrated HEC-HMS models are very sensitive to changes in the gridded precipitation, and the sensitivity is approximately equal for negative and positive changes in total rainfall. The HEC-HMS model sensitivity to precipitation ensures that the variability of the ensemble forecast precipitation members is preserved in the streamflow forecast ensemble.

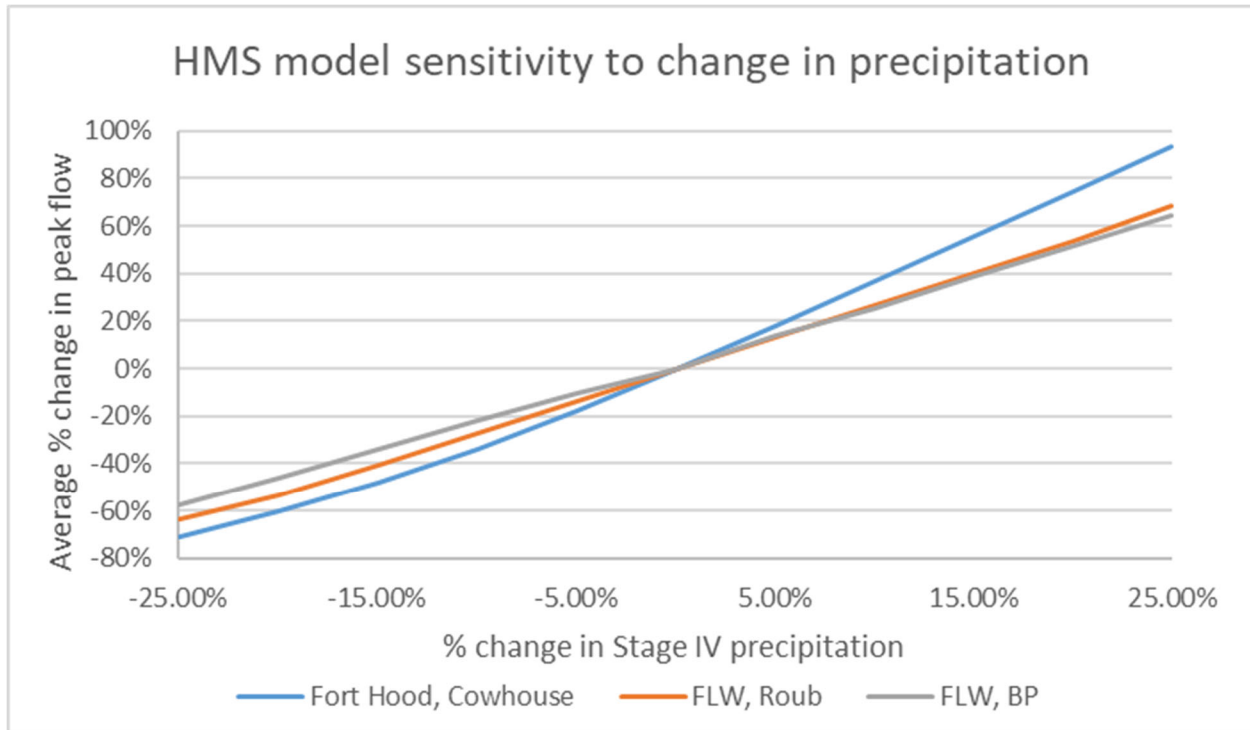


Figure 5.16: Average percentage change in the seven largest individual peak flow events during 2020-2021 for HEC-HMS model simulations due to changes in Stage IV gridded precipitation forcing. The Stage IV precipitation was modified uniformly +/- 25% at 5% increments. Events were run individually from hotstart files to isolate the effects of the changes in precipitation forcing on peak flow values.

5.3.4 Hydraulics Models

Hydro RF produces hydraulic parameters (e.g. flow depth, flow velocity, and water surface elevation) at specific locations with use of HEC-RAS. Steady, 1-D flow profiles were generated for each study reach and available gauge data is used to calibrate each model at each gauge location using flow rate and flow depth/water surface elevation data. The USGS periodically take field measurements at each gauge station to develop or update a rating curve. HEC-RAS rating curves were compared to recent USGS field measurements and published rating curves at all available locations (eg. Figure 5.17). HEC-RAS parameters (eg. channel Manning's n) were adjusted to minimize RMSE between the HEC-RAS and USGS rating curves. The calibrated models were used to develop rating curves (e.g. flow rate vs. flow depth, flow rate vs. flow velocity, flow rate vs. water surface elevation) at each critical infrastructure location.

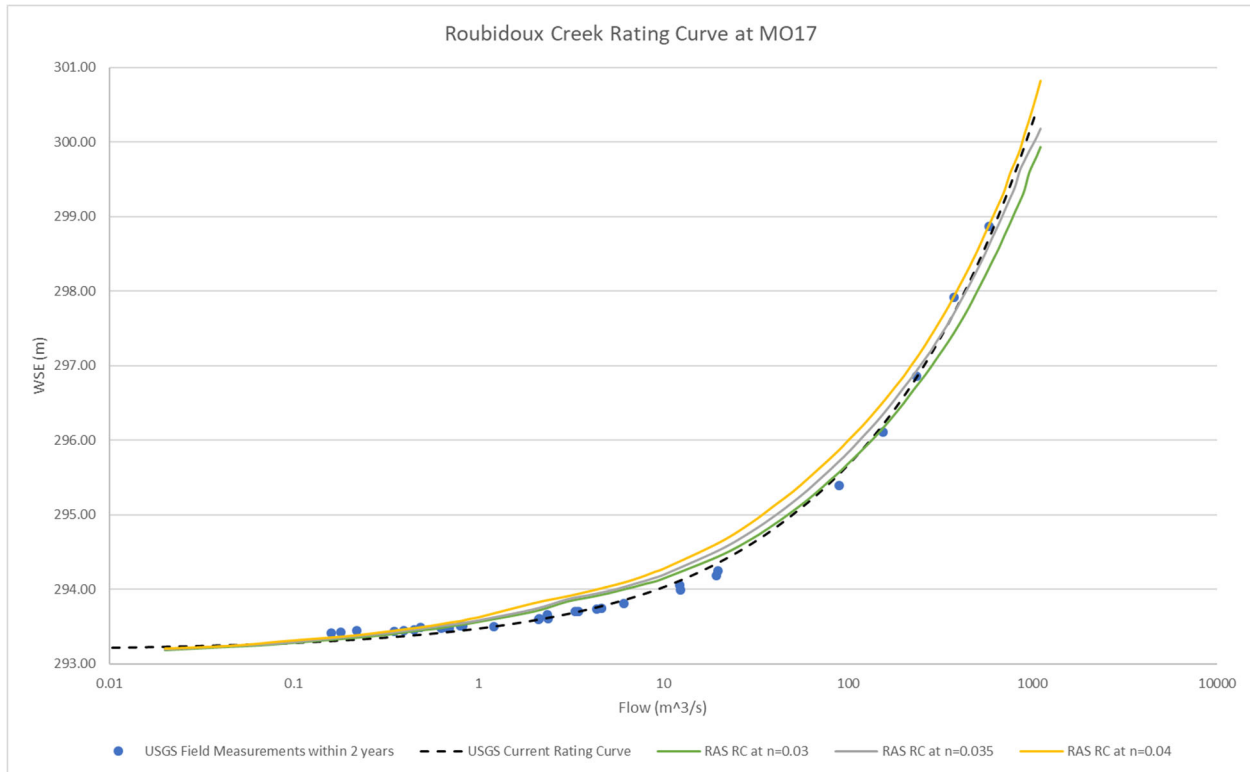


Figure 5.17: Rating curve of WSE vs flow at highway MO 17 over Roubidoux Creek. USGS rating curve field measurements within the last two years (blue points) and the published USGS rating curve at the time of RAS model calibration (black dashed line) are shown. RAS model rating curves for a range of channel Manning’s n values are shown (green, gray, yellow solid lines).

5.3.5 Downscaling Stress Test

HydroRF considers atmospheric forcing data from four sources, including the Global Forecast System (GFS), North American Mesoscale Forecast System CONUS nest (NAM-CONUS-nest), the HighRes Window (HiResW) products, and the Short Range Ensemble Forecast (SREF). While the NAM-CONUS-nest and HiResW products can be used in their native state, the SREF datasets must be downscaled to support Hydro RF applications. Information detailing the downscaling process is discussed thusly.

Hydro RF uses the WRF model to downscale coarser-resolution atmospheric forcing datasets to obtain sufficiently high-resolution mesoscale forecasts of precipitation. The WRF model is a physically-based Earth-system model that simulates atmospheric motion on a non-hydrostatic Eulerian grid. For the downscaling process, WRF uses the coarser model forecast datasets as initialization and boundary conditions. We configured the WRF version 4.3 (Skamarock et al. 2019) model to use three layers of nests, with the largest parent domain covering the western two-thirds of the CONUS and three sets of telescoping nests centered over the target installations, shown in Figure 5.18.

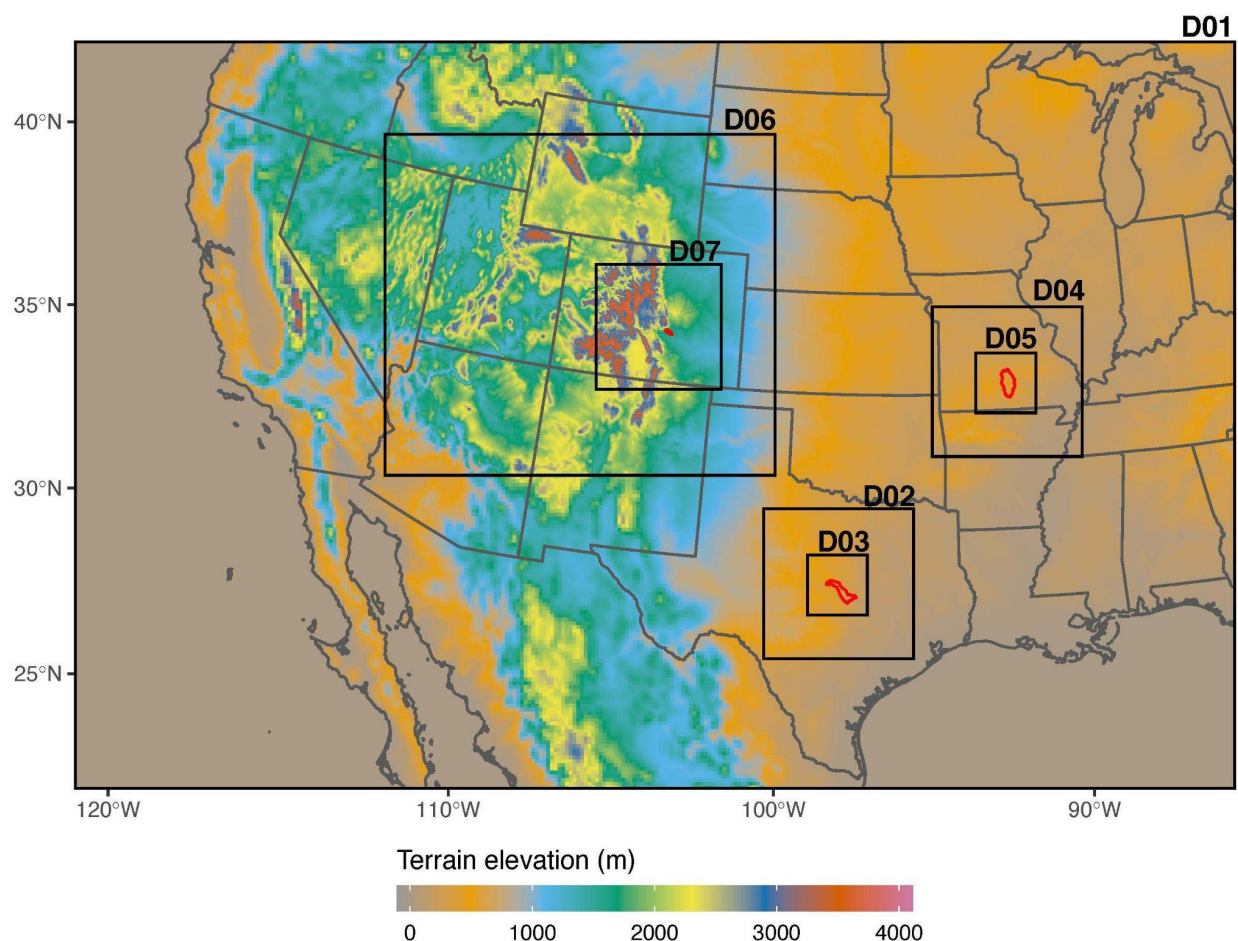


Figure 5.18: Telescoping WRF model domains with shading showing model terrain elevation in meters above sea level. The bright red outlines show the location of the catchments associated with the three installations.

The domain configuration gradually downscales the horizontal resolution of forecasts from 12 km in the first layer (domain 1 or D01), to 4 km in the second layer (D02, D04, and D06), down to 1.33 km in the third layer (D03, D05, and D07). We wish to briefly note the nests centered over Fort Carson, D06 and D07, are approximately three times as large as their counterparts for Fort Hood and FLW, D02-D05. This was purposely configured to better capture upstream weather systems and flow that transition over the nearby Rocky Mountains to the west of the Little Fountain and Rock Creek subbasins. The vertical level resolution is consistent for all seven domains and consists of 45 levels that are more tightly compressed near the surface and spread further out towards the model ceiling of 50 hPa.

To ensure a smooth transition from the native resolution of the input, we employed a warm start method, initializing WRF with the analysis and forecasts for 12 hours from the prior forecast cycle. For example, a 72-hour forecast cycle starting at 03Z(15Z) for the SREF or 00Z(12Z) for

the GFS would be preceded by twelve hours of spin up from the 15Z(03Z) SREF of 12Z(00Z) GFS earlier that day or the day before. This results in 84 total hours of WRF simulation time with the first 12 hours treated as spin-up for higher resolution phenomena and patterns to develop. These 12 hours are subsequently discarded and treated as independent of the 72 hours used in the Hydro RF ensemble.

The actual physics options used in our configuration of WRF were based on the physics used in the parent forecasts being downscaled. This was done to preserve physical processes being forecasted and the inherent spread among individual models and ensemble members. If a singular suite of physics were used the various forecasts being downscaled would essentially act only as differing initial conditions. Unfortunately readily available physics packages in WRF v4.3 limit our ability to mimic and preserve the parent physics. Certain physics options are only used by operational forecast centers and/or are not readily available for use in WRF without substantial user coding. As a result we are only able to downscale 10 of the 26 SREF members that utilize the WRF Advanced Research WRF (WRF-ARW) core and physics options listed in Table 5.8.

We downscaled and tested all 26 SREF members using a single configuration as part of our earlier stress test. The earlier stress test confirmed that a single WRF configuration approach inadvertently reduced the spread of members forced by downscaled SREF data.

Table 5.8: SREF ensemble members with equivalent physics options in WRFv4.3 and the physics packages used to downscale each member. Namelist options for each physics suite (in WRFv4.3) are included in brackets for convenience.

SREF Ensemble Member	Initial Conditions	Cumulus Scheme	Planetary Boundary Layer Scheme	Surface Layer Scheme	Microphysics Scheme	Radiation Scheme	Land Surface Model
WRF-SREF-arw_ctl	Rapid Refresh Model	Kain Fritsch (1)	Yonsei University (1)	Revised Mesoscale Model 5 (1)	WRF Single Moment 6 Class (6)	Rapid Radiative Transfer Model Global (4)	Noah (2)
WRF-SREF-arw_n1	Rapid Refresh Model	Betts Miller Janjic (2)	Mellor Yamada Janjic (2)	Mellor Yamada Janjic (2)	Ferrier (5)	Geophysical Fluid Dynamics Laboratory (99)	Noah (2)
WRF-SREF-arw_n2	Rapid Refresh Model	Kain Fritsch (1)	Yonsei University (1)	Revised Mesoscale Model 5 (1)	Ferrier (5)	Geophysical Fluid Dynamics Laboratory (99)	Noah (2)
WRF-SREF-arw_p2	Rapid Refresh Model	Betts Miller Janjic (2)	Mellor Yamada Janjic (2)	Mellor Yamada Janjic (2)	Thompson (8)	Rapid Radiative Transfer Model Global (4)	Noah (2)
WRF-SREF-arw_n3	Global Forecast System	Grell-Freitas (3)	Mellor Yamada Nakanishi and Niino (5)	Mellor Yamada Nakanishi and Niino (5)	Thompson (8)	Rapid Radiative Transfer Model Global (4)	Noah (2)
WRF-SREF-arw_n4	Global Forecast System	Betts Miller Janjic (2)	Mellor Yamada Janjic (2)	Mellor Yamada Janjic (2)	WRF Single Moment 6 Class (6)	Rapid Radiative Transfer Model Global (4)	Noah (2)
WRF-SREF-arw_n5	North American Data Assimilation System	Grell-Freitas (3)	Mellor Yamada Nakanishi and Niino (5)	Mellor Yamada Nakanishi and Niino (5)	Ferrier (5)	Geophysical Fluid Dynamics Laboratory (99)	Noah (2)
WRF-SREF-arw_p4	Global Forecast System	Kain Fritsch (1)	Yonsei University (1)	Revised Mesoscale Model 5 (1)	Ferrier (5)	Geophysical Fluid Dynamics Laboratory (99)	Noah (2)
WRF-SREF-arw_p5	North American Data Assimilation System	Kain Fritsch (1)	Yonsei University (1)	Revised Mesoscale Model 5 (1)	WRF Single Moment 6 Class (6)	Rapid Radiative Transfer Model Global (4)	Noah (2)
WRF-SREF-arw_p6	North American Data Assimilation	Grell-Freitas (3)	Mellor Yamada Nakanishi and Niino	Mellor Yamada Nakanishi and Niino	Thompson (8)	Rapid Radiative Transfer Model Global (4)	Noah (2)

	System		(5)	(5)			
--	--------	--	-----	-----	--	--	--

We tested 11 different ensemble configurations (Table 5.9) to assess the value of incorporating members forced by WRF-downscaled atmospheric datasets into our probabilistic Hydro RF ensemble settings. Ensemble 1 (Ens1) includes three members, all of which were forced by convective-resolving atmospheric models (NAM-CONUS nest and HiResW) that did not require additional downscaling through the WRF model. Ensemble configurations 2-6 (Ens2-Ens6) only include members forced by downscaled WRF datasets, starting with five members and increasing in number by five or six more members each instance. The remaining ensemble configurations (Ens7-Ens11) incorporate the three members from Ens1 into the WRF-downscaled configurations from Ens2-Ens6

Table 5.9: Ensemble configurations and their associated forecast length and number of ensemble members.

Atmospheric data source used to force the Hydro RF member	Ensemble configuration ID										
	Convective resolving members only	Downscaled members only					Downscaled and convective-resolving members				
	1	2	3	4	5	6	7	8	9	10	11
NAM-CONUS nest	✓						✓	✓	✓	✓	✓
HiResW-arw	✓						✓	✓	✓	✓	✓
HiResW-nmm/HiResW-fv3	✓						✓	✓	✓	✓	✓
WRF-GFS		✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
WRF-SREF-arw_ctl		✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
WRF-SREF-arw_n1		✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
WRF-SREF-arw_n3			✓	✓	✓	✓		✓	✓	✓	✓
WRF-SREF-arw_n5			✓	✓	✓	✓		✓	✓	✓	✓
WRF-SREF-arw_p5				✓	✓	✓			✓	✓	✓
WRF-SREF-arw_n2					✓	✓				✓	✓
WRF-SREF-arw_n4					✓	✓				✓	✓
WRF-SREF-arw_p2						✓					✓
WRF-SREF-arw_p4						✓					✓
WRF-SREF-arw_p6						✓					✓

Forecast length (hr)	48	72	72	72	72	72	48	48	48	48	48
Number of members	3	3	5	6	8	11	6	8	9	11	14

5.3.6 Risk Report communication strategie

Three prototype communication strategy figures were generated to report forecast flood hazard based on the HST and forecast confidence measured by the percentage of members forecasting hazardous conditions during each forecast window. All prototype figures were created for a 72-hour forecast. The first prototype figure is a tabular image, providing general information about the timing and confidence of the hazard forecast (Figure 5.19). The prototype tabular figure indicates the approximate level of risk (low, medium, and high, equating to green, amber, and red, respectively) based on member ensemble agreement of hazardous forecasts. The baseline confidence value is 50% of the ensemble members showing agreement, a high risk forecast. Medium (yellow) is ensemble members agreement of 40-49%. The confidence is provided as a color coded percentage of members forecasting hazardous conditions during each forecast interval. The forecast interval for the prototype is 12 hrs.

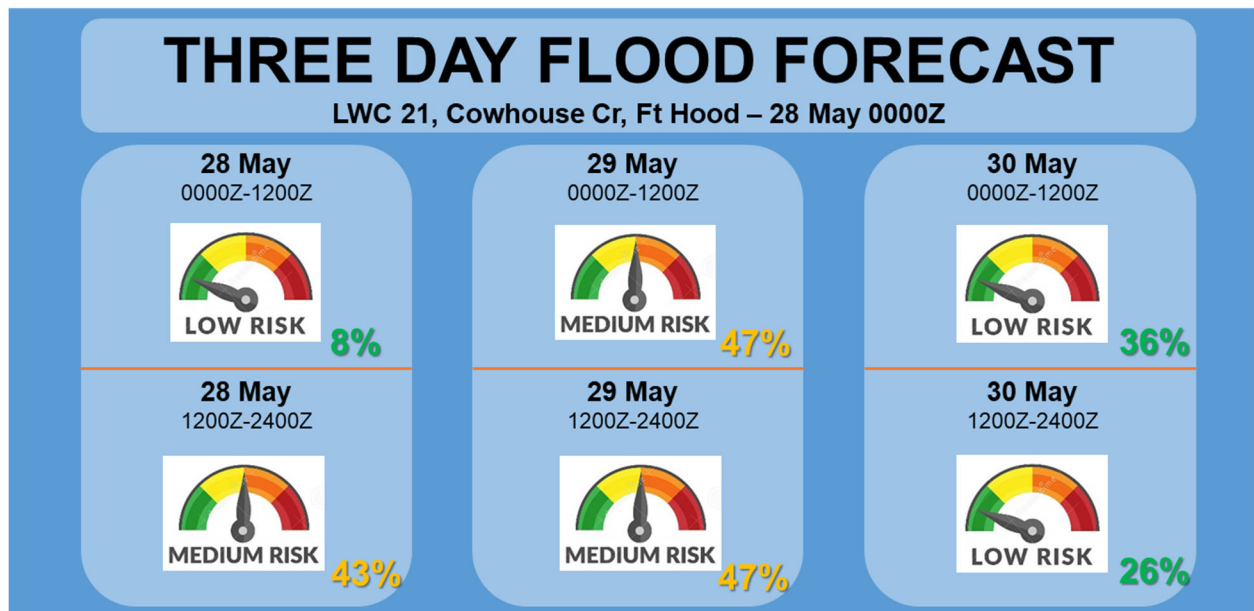


Figure 5.19: Tabular prototype risk communication figure, providing general information about the timing and confidence of the hazard forecast for LWC 21 at Fort Hood, initialized on 28 May 00Z.

The Heatmap prototype figure, Figure 5.20, better communicates the approximate hazard forecast timing and visualizes forecast confidence through model member agreement. The hourly HST value for each ensemble member is plotted on a graph of time vs HST. Showing all individual

ensemble members allows the decision maker to see when and to what extent the forecast members converge. The intensity of the background red/blue shows model agreement (confidence) with darker colors indicating more members above/below the $0.6 \text{ m}^2 \text{ s}^{-1}$ stability threshold (denoted by the yellow line). The stability threshold corresponding to the USGS observations is in black.

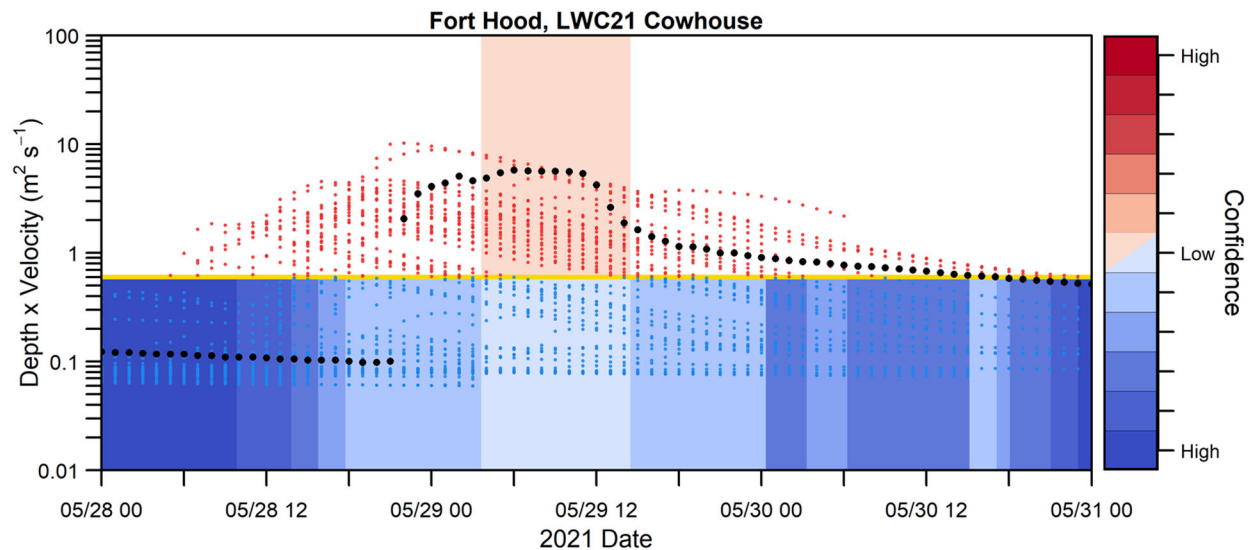


Figure 5.20: Heatmap prototype risk communication figure, providing approximate information about the timing and confidence of the hazard forecast for LWC 21 at Fort Hood, initialized on 28 May 00Z.

The third prototype communication figure, the Rating Curve, better communicates the reason for a forecasted hazardous conditions and indicates, generally, if the 72-hr forecast is a quiet period (Figure 5.21a) (little to no forecasted hazard) or an active period (Figure 5.21b) (some forecasted hazard). The hourly HST value for each ensemble member is plotted along the rating curve (relationship between flow depth and flow velocity) for a given infrastructure location. Each point to the right of the stability threshold (yellow line) is red to indicate a hazardous forecast. Bins are also populated along the x (water depth) and y (water velocity) axis to illustrate why a forecast is considered hazardous (excessive depth, excessive velocity, or both) and the portion of the hourly forecasts that fall into each bin. The Rating Curve figure is meant to supplement the first or second prototype figure.

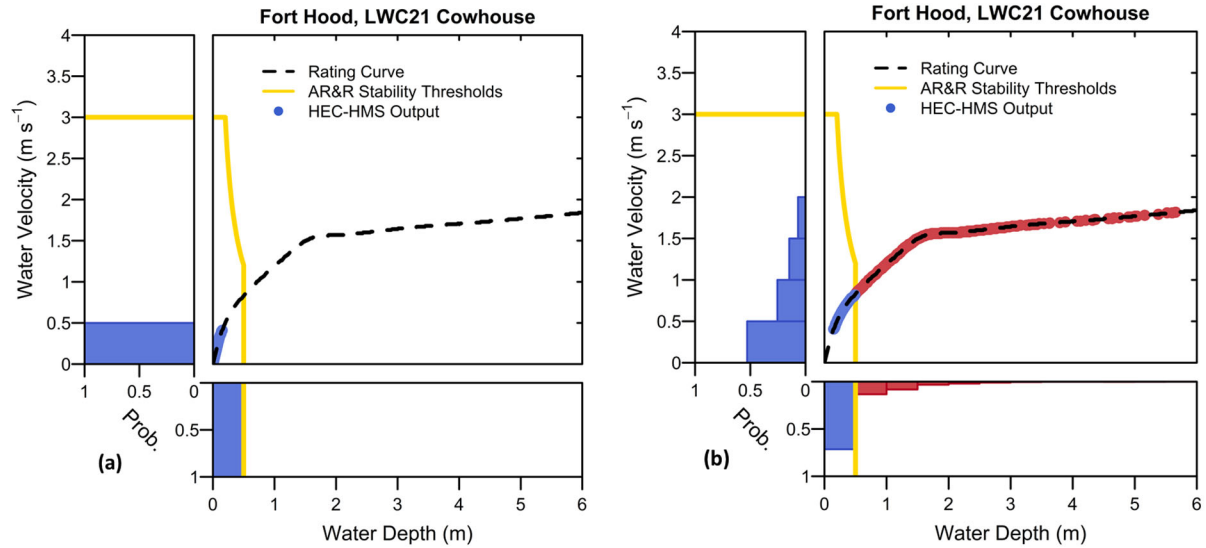


Figure 5.21: Rating Curve prototype risk communication figure, showing (a) a quiet forecast period for LWC 21 at Fort Hood, initialized on 10 March 00Z and (b) the reason for a hazardous 72 hr forecast outlook for LWC 21, initialized on 28 May 00Z. The latter figure indicates approximately 30% of all hourly forecast HST values show hazardous conditions at LWC 21 during the 72 hr forecast due to excessive forecasted water depth.

6.0 PERFORMANCE ASSESSMENT

6.1 Accurate probability forecasts for precipitation

Section Summary:

1. The skill of precipitation forecasts decreases as forecast lead times increase and neighborhood width decreases.
2. Precipitation forecast skill also decreases with increasing precipitation intensity.
3. Forecast performance is worse over smaller basins where only a few grid cells are available to characterize the rainfall or when the ensemble forecasts coincide with observed intense rainfall

The skill of precipitation forecasts from the Hydro RF atmospheric component is evaluated using FSS. Figure 6.1 shows FSSs for 1 hr accumulated precipitation $\geq 0.254 \text{ mm h}^{-1}$ as a function of forecast lead time for the three demonstration sites (Fort Hood, FLW, and Fort Carson). The skill scores are shown as means across the constituent NWP models at three neighborhood widths (0, 25, and 75 km). For all sites and all forecast lead times, FSSs are greater than zero, indicating the precipitation forecasts exhibit skill over the reference benchmark. FSSs are highest during the first 24 hours of the forecast period and gradually decrease through forecast hour 72, albeit with some hour-to-hour variability. At a neighborhood width of 25 km, day 1 (0–24 hr) FSSs are generally between 0.3 and 0.6 while day 3 (48–72 hr) FSSs are generally between 0.2 and 0.4. This decrease in skill with increasing lead time reflects the characteristic loss of predictability as forecasts extend further into the future, such that on average, day 1 forecasts of observed precipitation are more accurate than day 2 forecasts, and so on.

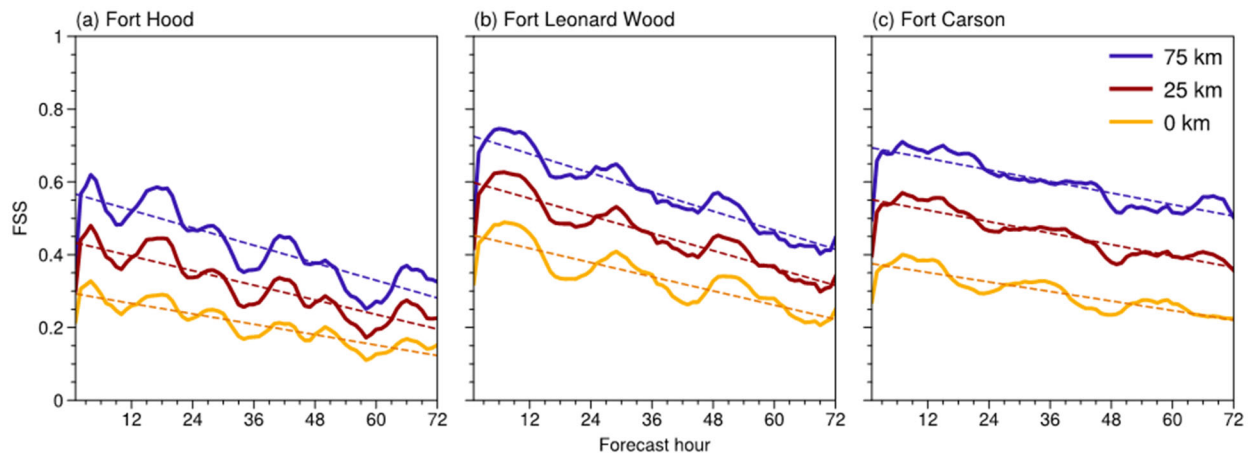


Figure 6.1: FSSs for (a) Fort Hood, (b) Fort Leonard Wood, and (c) Fort Carson as a function of forecast lead time and neighborhood width. Trends for each neighborhood width are shown by corresponding least squares regression lines (dashed).

Different levels of skill can be achieved when the neighborhood width used to compute FSSs is varied. The 0 km FSSs represent the grid-scale accuracy of the precipitation forecasts such that modeled and observed events must occur at the same hour and grid point. As discussed previously, high-resolution NWP models are inherently inaccurate at the grid scale, and FSSs are therefore lowest at the 0 km neighborhood width. The decreasing skill for decreasing neighborhood width suggests that forecasts for small catchments are likely to be less accurate, since small areas will be more sensitive to even modest discrepancies in the timing and placement of precipitation. On the other hand, increasing the neighborhood width to 75 km yields the best FSSs for the duration of the forecast period (i.e., > 0.5 for Fort Carson; > 0.4 for FLW; > 0.25 for Fort Hood). These results suggest that the NWP models are better in the approximate, rather than exact, placement of observed precipitation. For that reason, forecasts of precipitation over larger catchments are expected to outperform those of smaller catchments as there is a greater tolerance for timing and displacement errors.

Among the three sites, FLW and Fort Carson (Figure 6.1b–c) tend to have higher FSSs than Fort Hood (Figure 6.1a). The slightly lower skill for Fort Hood may reflect more difficult forecast situations during the verification seasons, or it may be an artifact of sample size (e.g., fewer precipitation events relative to the other sites). Nevertheless, the 25-km mean FSSs generally remain above 0.2 for the duration of the 3-day forecast period.

Since the Hydro RF atmospheric component is a conglomeration of different NWP models, skill can vary depending on the unique characteristics of each model. For example, Figure 6.2 compares FSSs across three forecast members—the HRRR, the 12 km NAM, and the 3 km NAM CONUS nest (NAMnest). During the first few forecast hours, the HRRR exhibits exceptionally high skill in 1 hr accumulated precipitation $\geq 0.254 \text{ mm h}^{-1}$, with FSSs as high as 0.8. At Fort Hood and FLW (Figure 6.2a–b), the HRRR outperforms the NAM and NAMnest through forecast hour 36, whereas at Fort Carson (Figure 6.2c), the HRRR skill approaches that of the NAMnest around forecast hour 18. Figure 6.2 demonstrates that each member of the Hydro RF atmospheric ensemble delivers, on average, a slightly different level of skill. However, model accuracy for individual forecast scenarios can vary significantly from the aggregate scores, motivating use of the full ensemble to capture the range of possible forecast outcomes for each event.

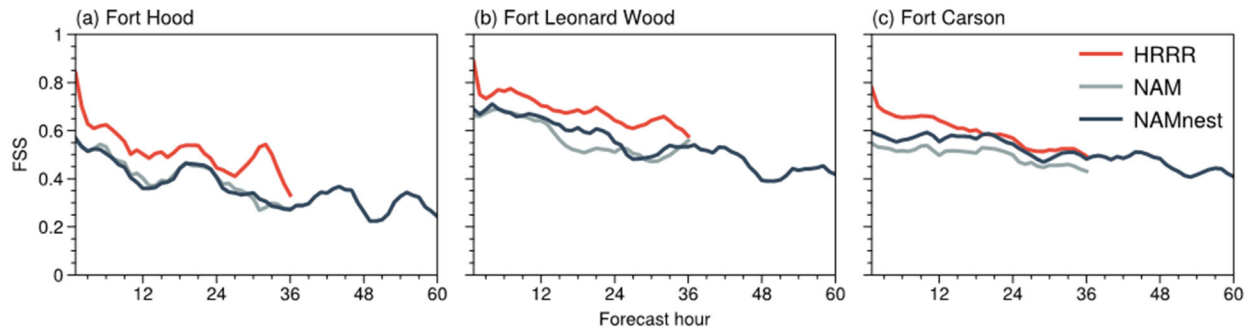


Figure 6.2: Comparison of FSSs for the 0.254 mm h⁻¹ threshold at (a) Fort Hood, (b) Fort Leonard Wood, and (c) Fort Carson for the HRRR, NAM, and NAMnest.

The impact of precipitation threshold on forecast skill as shown in Figure 6.3. At all three sites, FSSs decrease with increasing precipitation intensity. NWP models are better at predicting the occurrence and/or placement of light to moderate rainfall rates ($0.254\text{--}2.54\text{ mm h}^{-1}$) over moderate to heavy rates ($6.35\text{--}12.7\text{ mm h}^{-1}$). At higher rainfall rates, FSSs are only as high as ~ 0.2 in the first few forecast hours and decrease to less than 0.1 later in the forecast period.

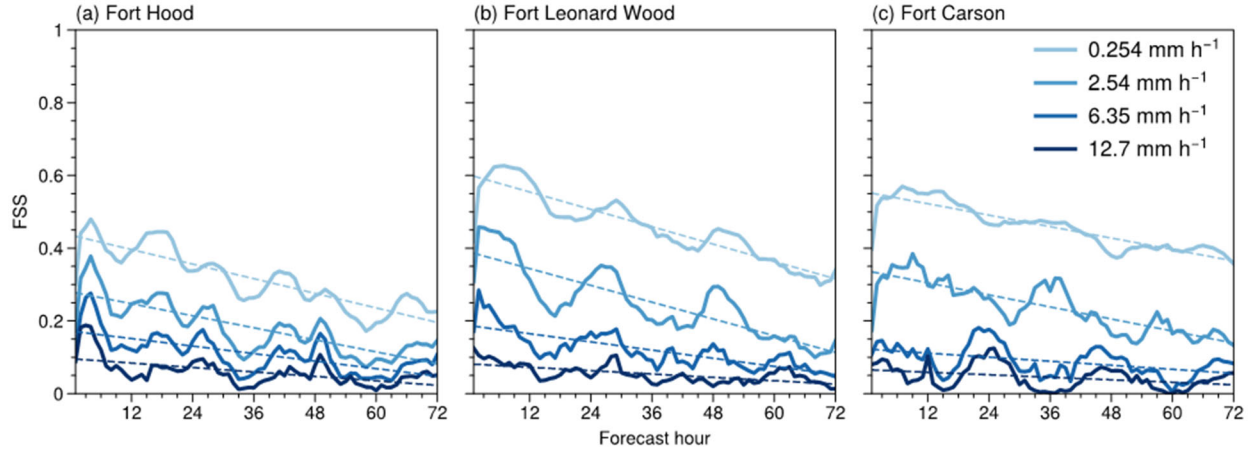


Figure 6.3: FSSs for (a) Fort Hood, (b) Fort Leonard Wood, and (c) Fort Carson as a function of forecast lead time and precipitation intensity. Trends for each precipitation intensity are shown by corresponding least squares regression lines (dashed).

Figure 4 shows 25 km FSSs aggregated by forecast day (i.e., 0–24 hr; 24–48 hr; 48–72 hr). As stated previously, forecast skill degrades with increasing lead time. For light precipitation (0.254 mm h^{-1}), FSSs decrease by 10–25% per day, whereas FSSs for moderate precipitation (6.35 mm h^{-1}) tend to decrease at slightly higher rates of 20–35% per day. The overall spread in model accuracy (shown by black bars and whiskers in Figure 6.4) tends to be largest for day 1 forecasts, which is at odds with the notion that ensemble spread is typically smallest at early lead times. The relatively large day 1 spread can be attributed to the aggregation of very high skill short-range forecasts (such as the HRRR) and lower-skill forecasts with spinup errors during the ~ 6 hr period after initialization. It is notable that the variability in skill for light precipitation at Fort Carson (Figure 6.4c) is small relative to Fort Hood and FLW (Figure 6.4a–b). This could be due to a higher frequency of diurnally and orographically driven convection in CO relative to TX and MO, which would act to constrain the variability in precipitation timing and placement. This behavior does not appear to hold at higher thresholds, as the Fort Carson ensemble variability at 6.35 mm h^{-1} (Fig. 4f) is more comparable to the other sites (Figure 6.4d–e).

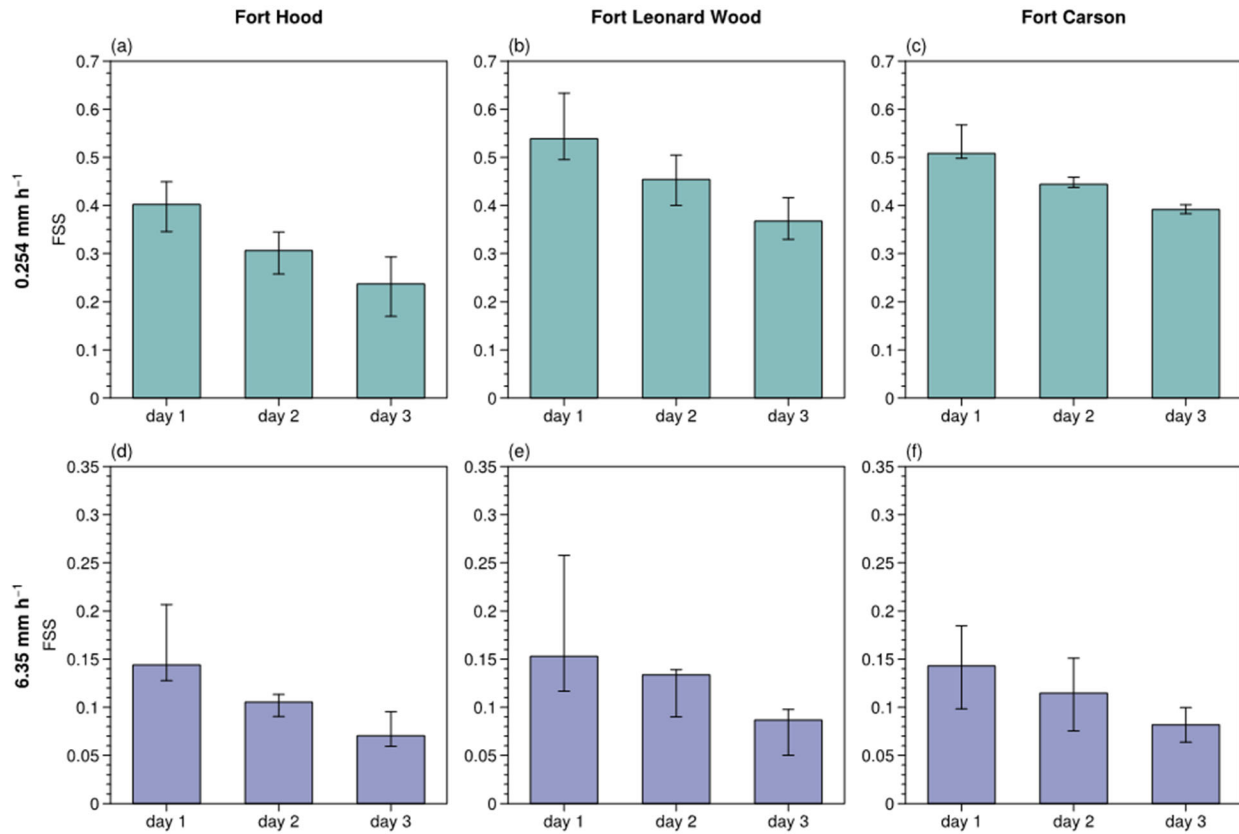


Figure 6.4: Median FSSs (filled bars) for the (a–c) 0.254 mm h⁻¹ and (d–f) 6.35 mm h⁻¹ precipitation thresholds at (left) Fort Hood, (center) Fort Leonard Wood, and (right) Fort Carson. The thin black bars and whiskers show the 10th and 90th percentiles in FSS across the ensemble members.

In addition to the statistical FSS analysis, numerous plots of cumulative area-averaged precipitation over time were created to visualize the accuracy of the forecast members compared to Stage IV precipitation. Area-average precipitation for Stage IV and all forecast members was created by clipping the datasets to basins and subbasins at each installation. By clipping the data to each basin we eliminate all precipitation falling outside, restricting the neighborhood to just the basins utilized by the HMS models. However, the timing of the basin-wide precipitation is preserved, allowing an hourly comparison of the forecast members to the Stage IV data. The following figures are representative of numerous precipitation events visualized at the three installations due to space restrictions.

Figure 6.5 shows the Stage IV and forecasted accumulated precipitation initialized at 27 May 2021, 00Z and 28 May 2021, 12Z for four of the study sub-basins at Fort Hood. The former forecast initialized approximately 42–48 hrs before a precipitation event and the latter initialized approximately 6–12 hrs before the same event. The forecast window corresponds to a widespread flash flood event. The forecast ensemble generally shows precipitation starting 6–12 hrs before the observed precipitation but underestimates total rainfall in the subsequent hours. As discussed

previously, the NWP models poorly predict high intensity precipitation. The May 2021 event at Fort Hood produced area average precipitation of 15-38 mm in 3-6 hours depending on the basin. The ensemble better predicts timing and quantity within the basins with less intense rainfall. For this event the ensemble performs better in the intermittent sized basins (House and Table Rock) than in the largest (Cowhouse) and smallest (Owl). The poor performance within the former basin is likely due to the inability of the ensemble to match the very high observed precipitation intensity while the poor performance within the latter basin is likely due to both the high observed precipitation intensity and the small size of the basin. The Owl Creek basin is approximately 20 km long and 8 km wide at its widest point.

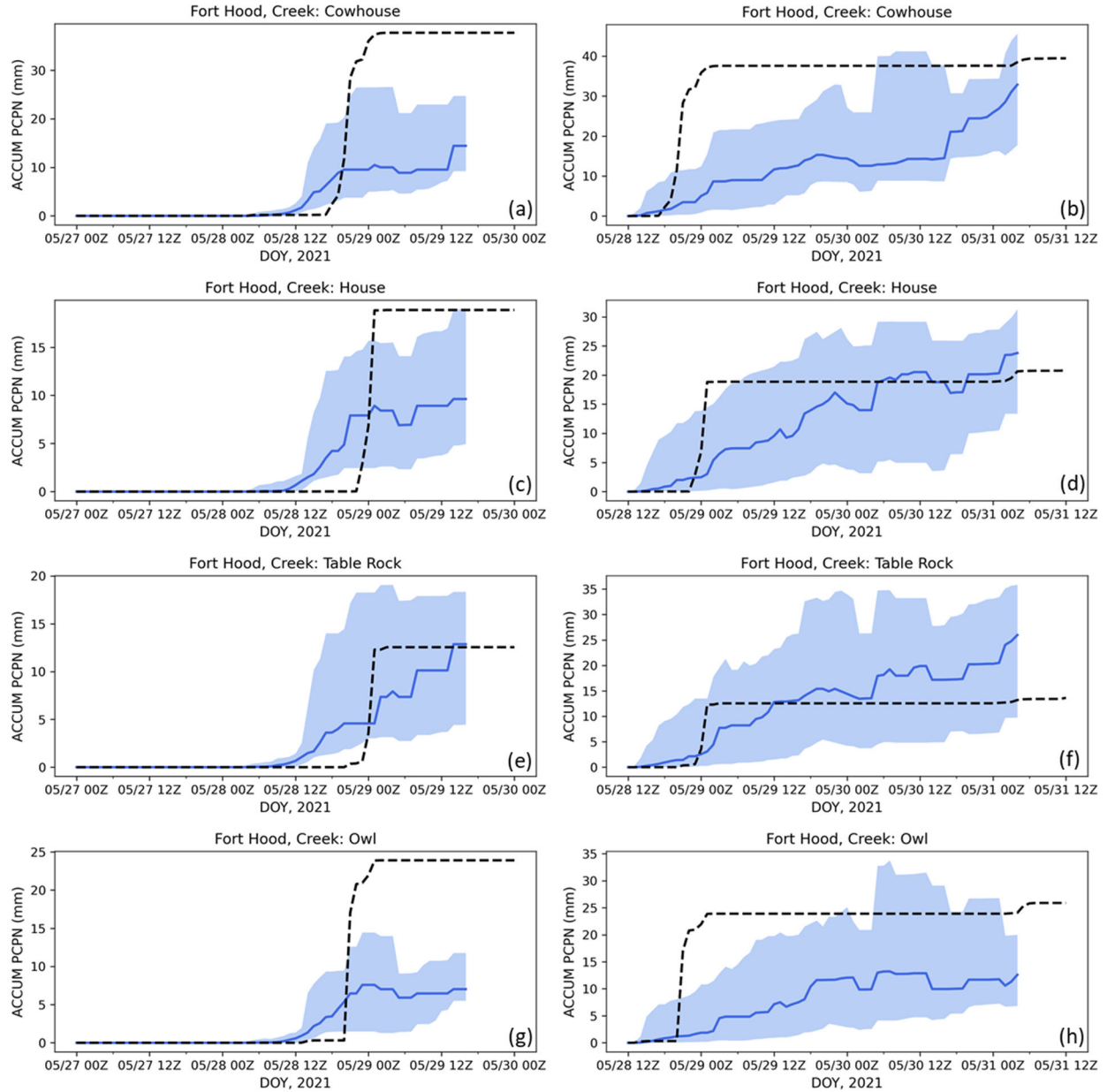


Figure 6.5: Accumulative area-average Stage IV (black dashed line) and median forecast member (blue line) precipitation at Fort Hood for a-b) Cowhouse Cr basin (1178 sq. km), c-d) House Cr basin (153 sq. km), e-f) Table Rock Cr basin (137 sq. km), and g-h) Owl Cr basin (53 sq. km). Forecasts initialized 27 May 2021 at 00Z and 28 May 2021 at 12Z. Blue shading extent represents the forecast ensemble 1st and 3rd quartiles.

Figure 6.6 shows the Stage IV and forecasted accumulated precipitation initialized at 9 March 2021, 12Z and 11 March 2021, 12Z for three of the study sub-basins at FLW. The forecast window corresponds with the largest flood event at FLW during 2021. The selected FLW event produced more precipitation than the May 2021 event at Fort Hood, but the intensity of the rainfall was less as it occurred over approximately 48 hrs. The first initialization, over 48 hrs before observed precipitation, shows early forecasted precipitation compared to observed, but comparable intensities and total rainfall. The later initialization, 6-12 hrs before observed precipitation starts,

again shows comparable rainfall intensity and total, especially over the larger Big Piney basin during the first 48 hrs of the forecast. The ensemble forecast over the smallest basin (Turnbull Hollow, a tributary to Roubidoux Cr) performs poorly compared to the larger basins, underestimating total rainfall by almost half the observed. The poor performance is likely due to the small basin size (15.5 sq. km) which is two orders of magnitude smaller than the Big Piney basin and 33 times smaller than the Roubidoux basin. The Turnbull Hollow basin is approximately 5 km long and 3 km wide, meaning parts of only three or four precipitation grid cells cover the basin.

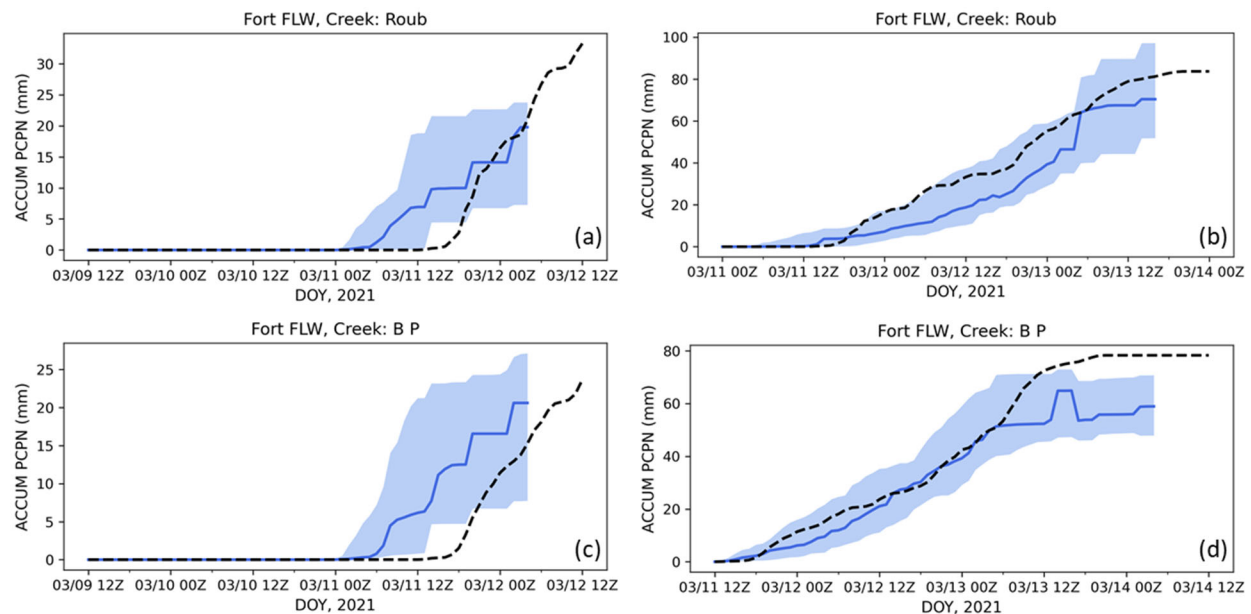


Figure 6.6: Accumulative area-average Stage IV (black dashed line) and median forecast member (blue line) precipitation at Fort Leonard Wood for a-b) Roubidoux Cr basin (519 sq. km), c-d) Big Piney River basin (1536 sq. km), and e-f) Turnbull Hollow basin (15.5 sq. km). Forecasts initialized 9 March 2021 at 12Z and 11 March 2021 at 12Z. Blue shading extent represents the forecast ensemble 1st and 3rd quartiles.

Figure 6.7 shows the Stage IV and forecasted accumulated precipitation initialized at 29 June 2021, 00Z and 30 June 2021, 12Z for the Little Fountain Creek study basin at Fort Carson. The ensemble forecast over the Little Fountain basin is similar to the forecast over Turnbull basin at FLW. The first initialization, over 48 hrs before observed precipitation, shows early forecasted precipitation compared to observed. The later initialization, 12 hrs before observed precipitation starts, performs poorly, underestimating total rainfall by almost half the observed. The poor performance is again likely due to the small basin size (69 sq. km) and shape (approximately 24 km long and 6 km wide). The high intensity observed rainfall, which occurred between 05Z-07Z on 1 July 2021 is also not captured by the ensemble forecast.

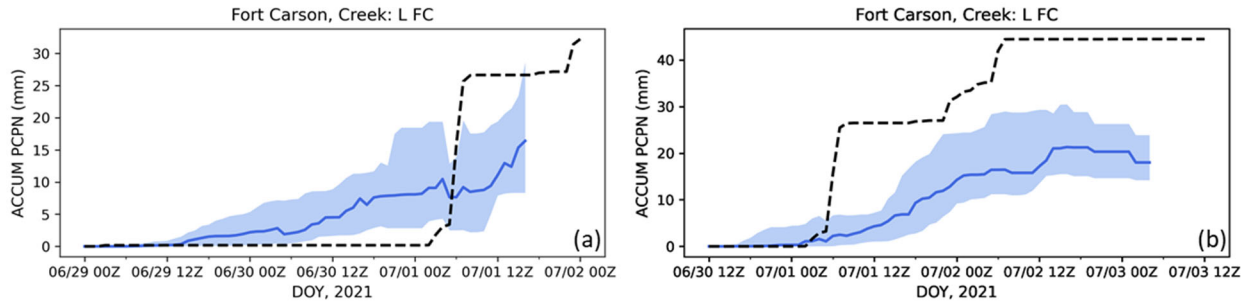


Figure 6.7: Accumulative area-average Stage IV (black dashed line) and median forecast member (blue line) precipitation at Fort Carson for Little Fountain Cr basin (69 sq. km). Forecasts initialized 29 June 2021 at 00Z and 30 June 2021 at 0000Z. Blue shading extent represents the forecast ensemble 1st and 3rd quartiles.

The ensemble forecasts perform well ($FSS > 0.4$) over larger areas/basins, especially for less intense rainfall. Performance is worse over smaller basins where only a few grid cells are available to characterize the rainfall or when the ensemble forecasts coincide with observed intense rainfall. Areas of localized heavy rainfall are often the product of storm-scale circulations still unresolved by models with convection-allowing resolutions. Challenges remain in predicting particularly intense rainfall rates. However, increased forecast skill at light to moderate thresholds gives confidence in the ensemble's ability to place the overall precipitation system correctly, despite uncertainty in the details of embedded heavy precipitation.

6.2 Accurate probability forecasts for streamflow time series

Section Summary:

1. The performance of the forecast streamflow ensemble reflects the shortcomings of both the forecast precipitation ensemble and the calibrated HEC-HMS model concerning the timing and quantity of forecast flow.
2. The BSS results indicate the full ensemble is skillful at the majority of sites, though we note, however, that flood hazards are not a common occurrence, so the BSS may be biased by the accurate prediction of the null events.

A review of HEC-HMS model simulation performance using Stage IV forcing during the 2020-2021 Hydro RF validation periods is provided to contextualize the ensemble forecast streamflow and hazard assessments. The HEC-HMS models were forced with Stage IV precipitation and NLDAS gridded ET for continuous simulations from 1 Jan 2020 to 20 July 2021. Model performance is assessed during the Hydro RF validation periods (01 Apr 2020 - 15 June 2020 and 01 Mar 2021 - 16 July 2021). Additional gauge locations are available for the 2020-2021 HEC-HMS performance assessment due to the HOBO sensors deployed for this project at Fort Hood and FLW and additional USGS gauge stations that were installed within Fort Hood on Cowhouse Creek and House Creek in 2018-2019.

Model performance along Cowhouse Creek in 2020 (Figure 6.8a) and 2021 (Figure 6.8b) underestimates peak flow during larger events and overestimates flow during smaller events.

Statistically model performance increases downstream (Table 6.1), as the simulated peak flows for the larger events are more underestimated for the upstream gauge (Pidcoke) than the downstream gauge (West Range Rd) and overall performance is better during 2021 than during 2020. Model performance for the basins not directly calibrated (House Creek, Figure 6.8c-d; Table Rock Creek, Figure 6.9a-b, and Owl Creek, Figure 6.9c-d) underestimates peak flow, especially along House and Owl Creeks. Peak flows along these reaches are 25%-50% of the observed peaks. Underperformance along these tributary reaches is also reflected statistically (Table 6.1).

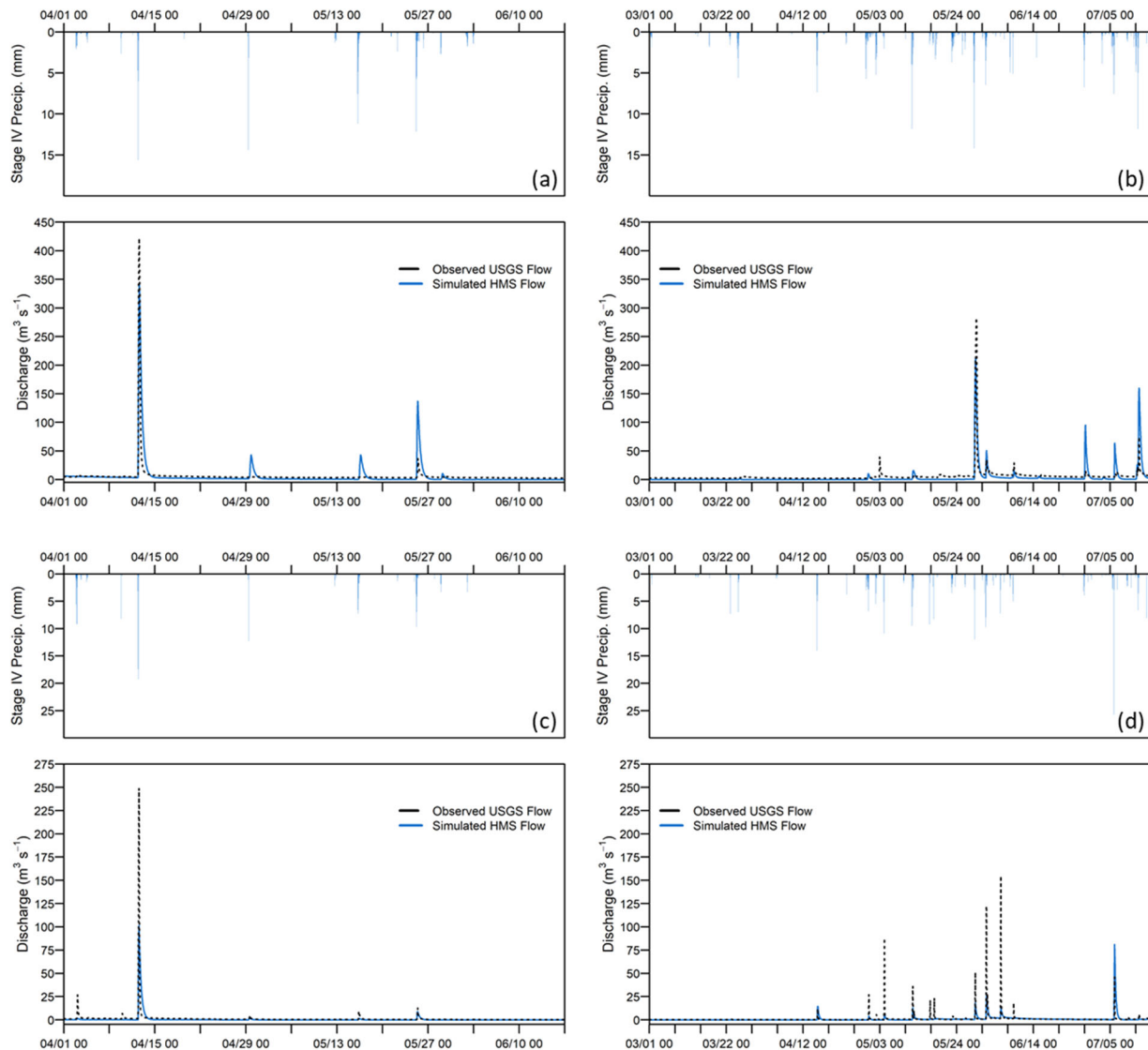


Figure 6.8: Graphs of calibrated HEC-HMS performance during the 2020 (left) and 2021 (right) validation periods for Cowhouse Creek (a-b) at West Range Rd and for House Creek (c-d) at West Range Rd. a) The top plot is the hyetograph of hourly Stage IV precipitation averaged across the catchment upstream of the stream gauge. The bottom plot shows the

observed streamflow (black dashed line) from the USGS or HOBO gauge and the calibrated (blue line) HEC-HMS modeled streamflow from long term simulations forced by the Stage IV precipitation forced in the top plot. (b-d) as in (a), but for other timeframes or locations.

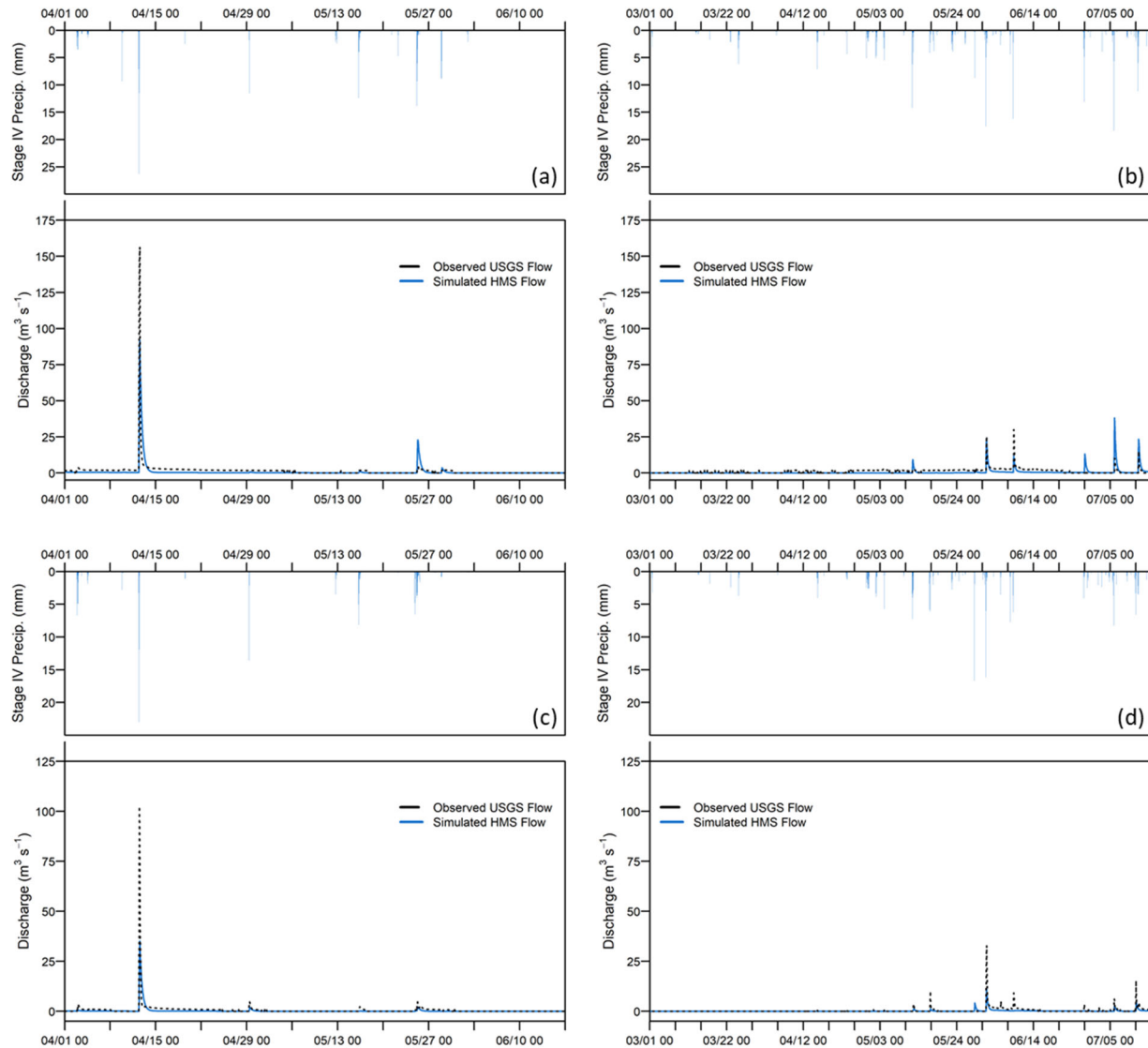


Figure 6.9: Graphs of calibrated HEC-HMS performance during the 2020 (left) and 2021 (right) validation periods for Table Rock Creek (a-b) at Antelope Rd and for Owl Creek (c-d) at LWC 26. a) The top plot is the hyetograph of hourly Stage IV precipitation averaged across the catchment upstream of the stream gauge. The bottom plot shows the observed streamflow (black dashed line) from the USGS or HOBO gauge and the calibrated (blue line) HEC-HMS modeled streamflow from long term simulations forced by the Stage IV precipitation forced in the top plot. (b-d) as in (a), but for other timeframes or locations.

Calibrated HEC-HMS model performance along Roubidoux Creek in 2020 (Figure 6.10a) and 2021 (Figure 6.10b) is mixed, underestimating some larger peak flow events and closely matching other events. The Big Piney River simulation (Figure 6.10c-d) is similar, though the simulation

falling limbs fall too quickly and the baseflow is often too low. Statistically model performance is comparable between the Roubidoux and Big Piney and between gauge locations along each reach (Table 6.1). Model performance for the basin not directly calibrated (Turnbull Hollow, Figure 6.10) significantly underestimates peak flow, showing almost no hydrologic response during 2021. The Turnbull Hollow subbasin is the smallest gauged subbasin with demonstration sites. Turnbull Hollow remained ungauged in 2020 due to deployment delays caused by COVID 19. Turnbull Hollow underperformance is also reflected statistically (Table 6.1).

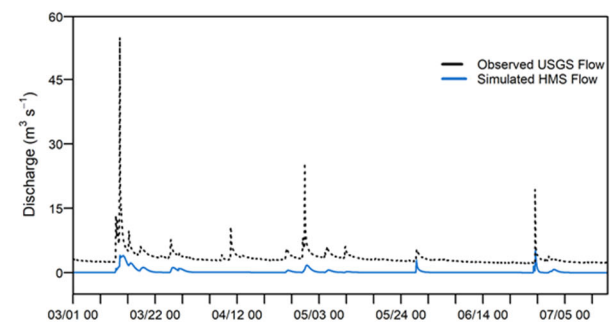
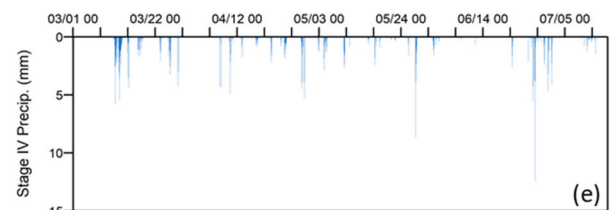
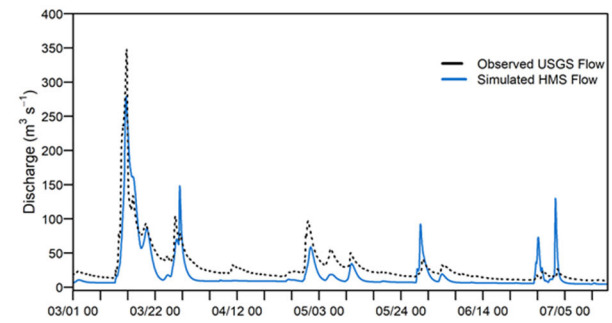
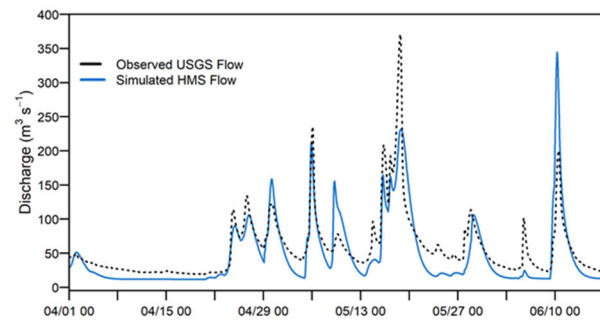
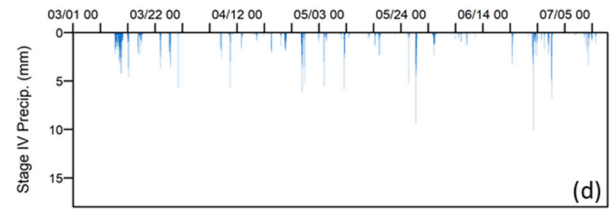
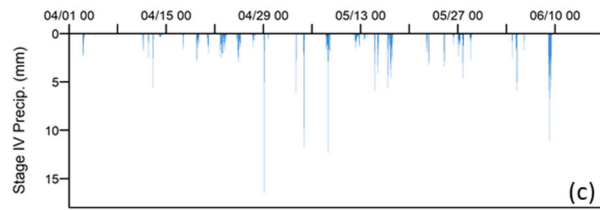
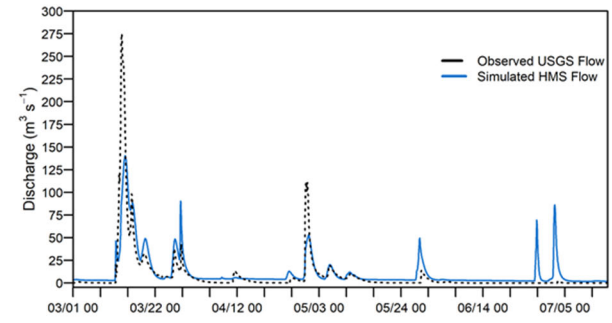
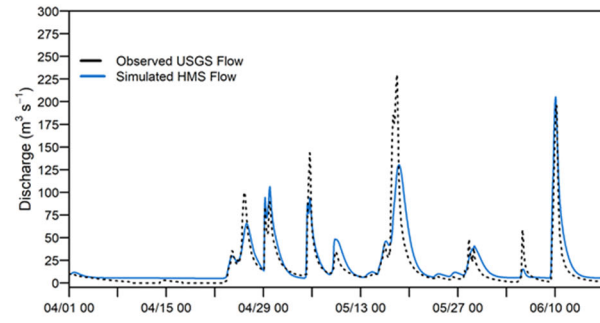
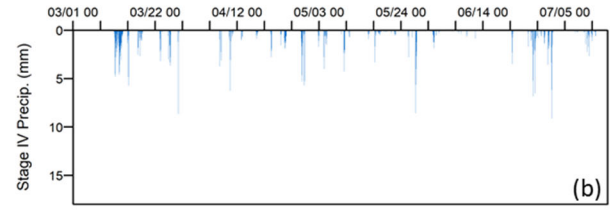
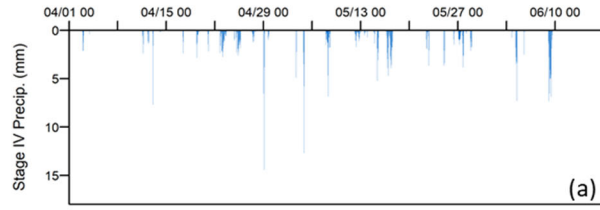


Figure 6.10: Graphs of calibrated HEC-HMS performance during the 2020 (left) and 2021 (right) validation periods for Roubidoux Creek (a-b) at Polla Rd, for the Big Piney River (c-d) at East Gate Rd, and for Turnbull Hollow (e) at FLW U. a) The top plot is the hyetograph of hourly Stage IV precipitation averaged across the catchment upstream of the stream gauge. The bottom plot shows the observed streamflow (black dashed line) from the USGS or HOBO gauge and the calibrated (blue line) HEC-HMS modeled streamflow from long term simulations forced by the Stage IV precipitation forced in the top plot. (b-e) as in (a), but for other time frames or locations.

Little Fountain Creek at Fort Carson experienced no measurable streamflow during 2020 and two small events in July of 2021 (Figure 6.11a-b). Both July 2021 events are overestimated by the HEC-HMS simulation but the small number of datapoints prevents analysis of general model performance. The performance statistics are presented in Table 6.1, but are not meaningful due to the lack of data points.

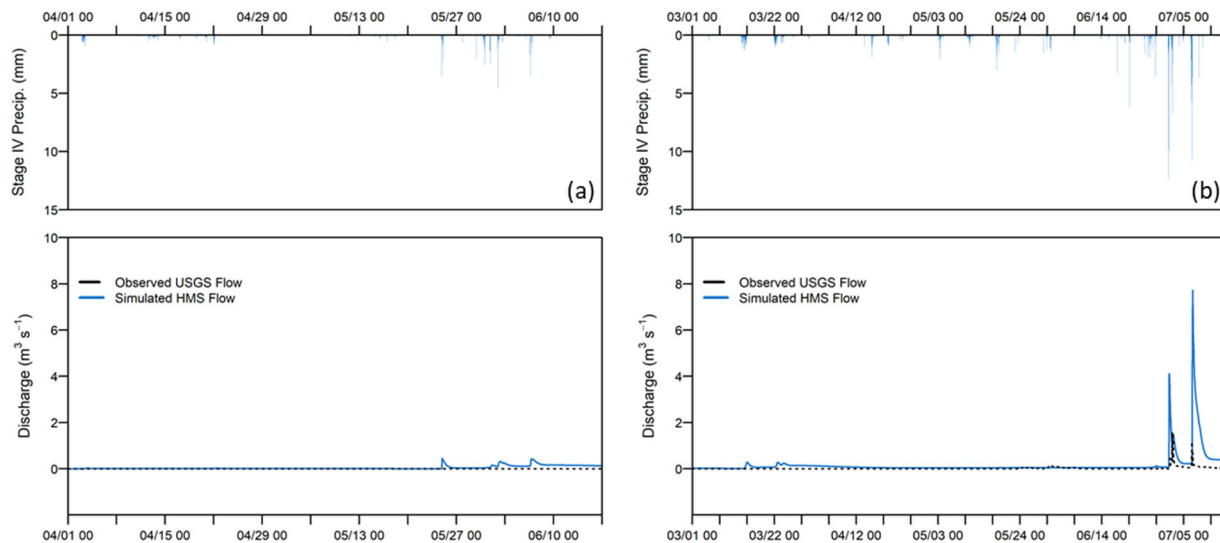


Figure 6.11: Graphs of calibrated HEC-HMS performance during the 2020 (left) and 2021 (right) validation periods for Little Fountain Creek (a-b) at R5. a) The top plot is the hyetograph of hourly Stage IV precipitation averaged across the catchment upstream of the stream gauge. The bottom plot shows the observed streamflow (black dashed line) from the USGS or HOBO gauge and the calibrated (blue line) HEC-HMS modeled streamflow from long term simulations forced by the Stage IV precipitation forced in the top plot. (b) as in (a), but for another timeframe.

Table 6.1: Model evaluation statistics for the calibrated Fort Hood, FLW, and Fort Carson HEC-HMS performance at all available USGS and HOBO gauge locations during the 2020 and 2021 Hydro RE validation periods.

Installation	Stream	Location	Drainage Area (km ²)	Validation Period	NSE	KGE	RSR	RMSE (m ³ /s)
Fort Hood	Cowhouse Cr	West Range Rd	1442.6	2020	0.65	0.80	0.60	12.80
				2021	0.51	0.55	0.70	10.78
		Georgetown Rd	1388.2	2020	0.03	0.30	0.98	14.40
				2021	0.46	0.57	0.73	9.58
		Pidcoke *	1178.4	2020	-0.06	-0.21	1.03	8.86
				2021	0.70	0.81	0.55	8.89
	House Cr	West Range Rd	153.1	2020	-0.32	-0.12	1.15	10.35
				2021	-0.38	0.02	1.17	6.38
		Georgetown Rd	32.9	2020	-0.13	0.38	1.06	2.22
				2021	-1.58	0.07	1.61	2.46
	Table Rock Cr	Antelope Rd	136.9	2020	0.13	0.34	0.93	5.89
				2021	-1.08	0.21	1.44	2.39
	Owl Cr	LWC 26	53.1	2020	-0.17	-0.14	1.08	3.97
				2021	0.07	-0.03	0.96	1.30
FLW	Roubidoux Cr	Polla Rd	707.1	2020	0.74	0.73	0.51	16.13
				2021	0.59	0.52	0.64	17.31
		FLW T	519.4	2021	0.31	0.42	0.83	19.83
				2020	0.53	0.63	0.68	15.70
		MO 17 *	455.0	2021	0.38	0.50	0.79	12.87
	Big Piney River	East Gate Rd	1535.9	2020	0.64	0.76	0.60	28.53
				2021	0.62	0.61	0.62	21.32
		Western Rd *	1450.4	2020	0.60	0.78	0.63	28.16
				2021	0.58	0.67	0.65	20.43
	Turnbull Hollow	FLW U	15.5	2021	-2.39	-0.22	1.84	3.61
Fort Carson	Little Fountain Creek	R5 *	69.4	2021	-119.72	-15.24	10.99	0.83

* denotes locations with longterm USGS stream gauges used for HMS model calibration

The Fort Hood and FLW simulations habitually underestimate observed peak flow for the largest flow events in 2020-2021, partially because we calibrated the HMS models with statistics based on a continuum of streamflow as opposed to other techniques that prioritize events like calibrating to peaks. In the future, the calibration process should potentially more heavily reward peak flow agreement and calibration periods with a large number of large flow events should be

prioritized. The underestimation of peak flows becomes more pronounced as the subbasins size decreases. RSR results (excluding Fort Carson) compared to basin size shows that error increases for smaller basins (Figure 6.12). The Stage IV precipitation forcing may also contribute to peak flow underestimation in the small subbasins. The 4 km grid scale of Stage IV may not capture very intense, localized rainfall, especially in regions where precipitation is driven by strong convective storms. Localized intense rainfall becomes the predominant driver of flood events as basin size becomes smaller, therefore, the model simulations may be constrained by both the model calibration and the gridded precipitation forcing.

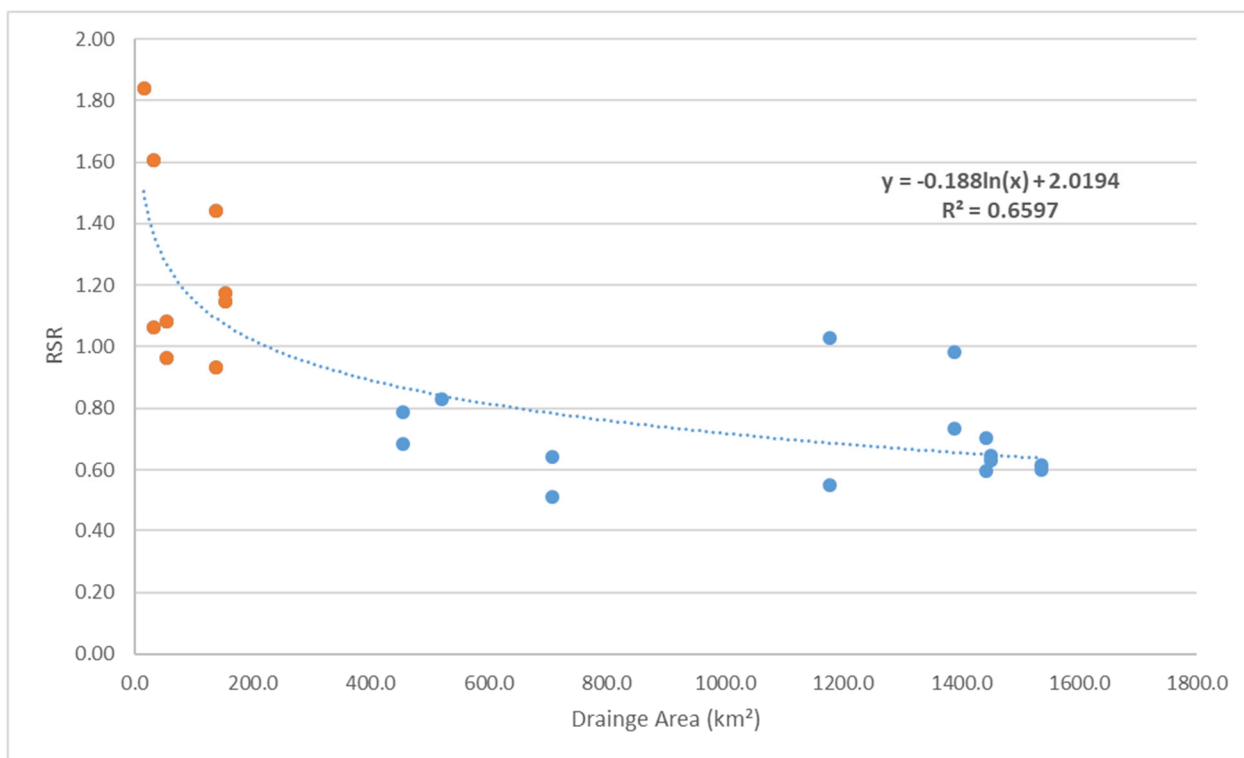


Figure 6.12: RSR performance for the calibrated Fort Hood, FLW, and Fort Carson HEC-HMS models at all available USGS and HOBO gauge locations during the 2020 and 2021 Hydro RF validation periods vs. gauge location drainage basin size. Gauges located along the same stream reach as the long term USGS gauges used for HMS model calibration/validation appear in blue. Gauge locations within uncalibrated sub-basins appear in orange.

We test if the full ensemble streamflow forecast is skillful over the assumption that present conditions (hazard determination at forecast initialization) are always persistent (never change over the 72 hr forecast window). Table 6.2 provides the BSS scores at available locations at Fort Hood and FLW, some locations are not included due to data constraints. The results indicate the full ensemble is skillful at the majority of sites, though we note, however, that flood hazards are not a common occurrence, so the BSS may be biased by the accurate prediction of the null events.

Table 6.2: Model evaluation statistics for the calibrated Fort Hood, FLW, and Fort Carson HEC-HMS performance at all available USGS and HOBO gauge locations during the 2020 and 2021 Hydro RF validation periods.

	SiteName	BSS
Hood	LWC03_House	0.55
	LWC06_House	0.48
	LWC07_House	0.35
	Antelope_TableRock	-0.48
	LWC09_TableRock	0.44
	LWC10_TableRock	0.49
	LWC11_TableRock	0.64
	LWC12_TableRock	0.60
	LWC13_TableRock	-0.25
	LWC14_TableRock	0.16
	LWC15_TableRock	-0.18
	LWC16_TableRock	0.16
	LWC18_Cowhouse	0.04
	LWC19_Cowhouse	-1.80
	LWC20_Cowhouse	-0.23
	LWC21_Cowhouse	0.05
	LWC22_Cowhouse	0.07
	LWC26_Owl	0.41
FLW	FLW8_Roub	0.03
	SpringfieldRd_DS_Roub	-0.72
	StillwaterRd_Roub	-0.78
	FLWU_Turnbull	0.05

In addition to the statistical BSS analysis, numerous plots of ensemble streamflow were created to visualize the accuracy of the forecast members compared to observed and simulated streamflow. The following figures are representative of numerous flood events visualized at the three installations due to space restrictions.

Figure 6.13 shows the Stage IV and forecast precipitation hyetographs and observed, simulated, and forecast hydrographs initialized at 28 May 2021, 00Z or 31 May 2021, 12Z for four of the study sub-basins at Fort Hood. The forecast window corresponds to a widespread flash flood event. The precipitation forecast ensemble generally shows precipitation starting 6-12 hrs before or following the observed precipitation. The resulting streamflow ensemble agreement (darker green shade) generally underestimates the observed streamflow and is either 3-6 hrs early or late, however, a minority of members match or overestimate the observed peak flows. The general underestimation of ensemble peak flow is at least partially attributable to the HEC-HMS model

given the persistent underestimation of the calibrated simulations, but the timing of the ensemble hydrographs is due to the missed timing of the forecast precipitation members.

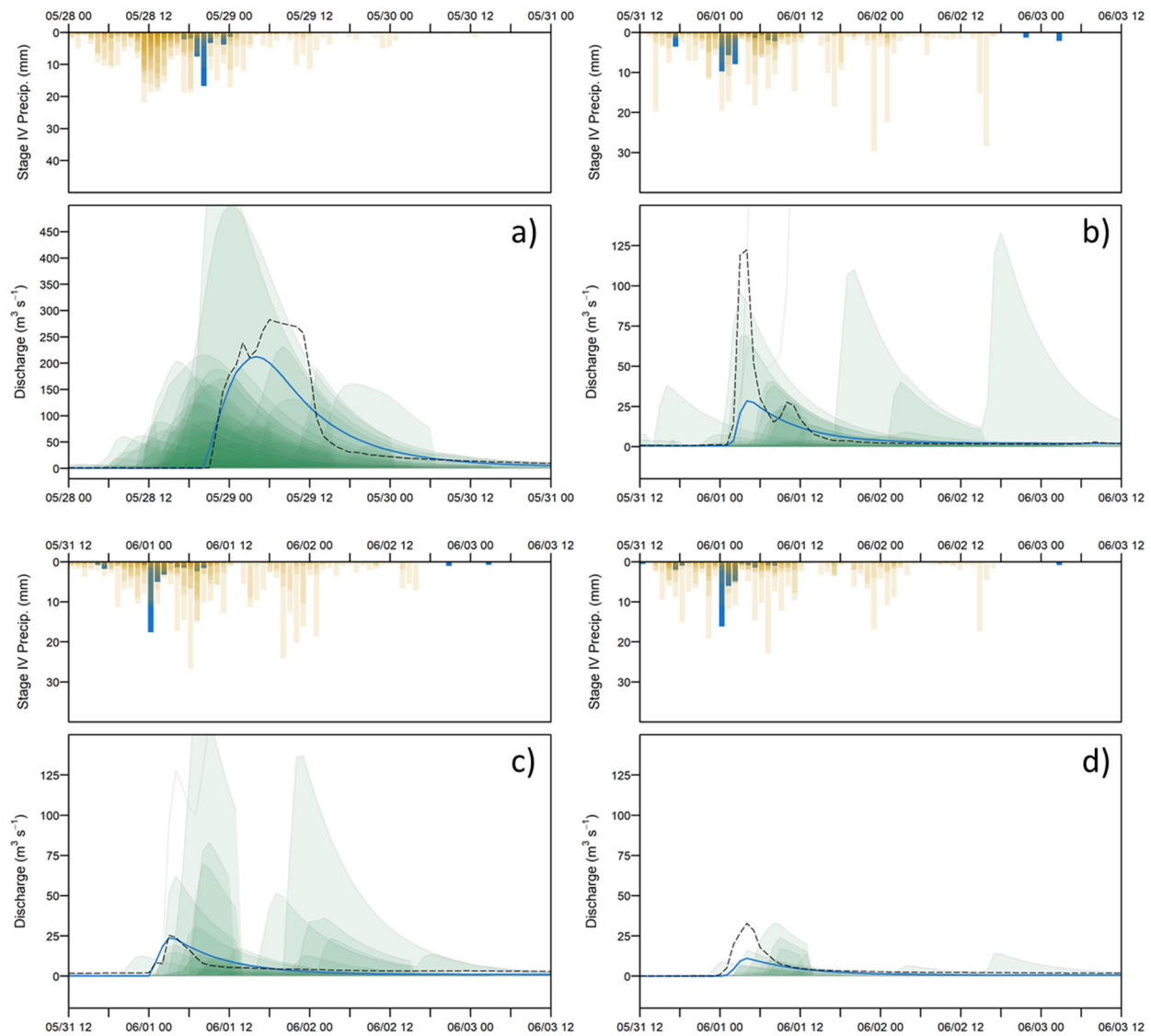


Figure 6.13: a) Ensemble forecast streamflow performance at Fort Hood LWC 21 initialized at 28 May 2021, 00Z. The top plot is the hyetograph of hourly Stage IV precipitation (blue) and individual ensemble precipitation (gold) averaged across the area of the Cowhouse Creek catchment upstream of LWC 21. The bottom plot shows the observed streamflow (black dashed line), the simulation forced by Stage IV precipitation (blue line) and each green shading represents a streamflow time series corresponding to one of the ensemble forecast members. (b)-(d) as in (a), but for LWC 7 on House Creek, LWC 13 on Table Rock Creek, and LWC 26 on Owl Creek, respectively, all initialized at 31 May 2021, 12Z.

Figure 6.14 shows the Stage IV and forecast precipitation hyetographs and observed, simulated, and forecast hydrographs initialized at 13 March 2021, 12Z for three of the study sub-basins at

FLW. The forecast window corresponds to a widespread flood event. The precipitation forecast ensemble shows precipitation starting early or approximately on time compared to the observed precipitation. The resulting streamflow ensemble agreement (darker green shade) generally underestimates the observed streamflow, especially for the Roubidoux and Turnbull hollow, which should be expected given the persistent underestimation of the calibrated simulations. The timing of the hydrograph ensemble along these reaches is good. Visually, the ensemble agreement outperforms the Stage IV simulation along the Roubidoux. The Big Piney ensemble more closely matches the observed peak flow but is early in arrival time, again, due to the early timing of the precipitation ensemble.

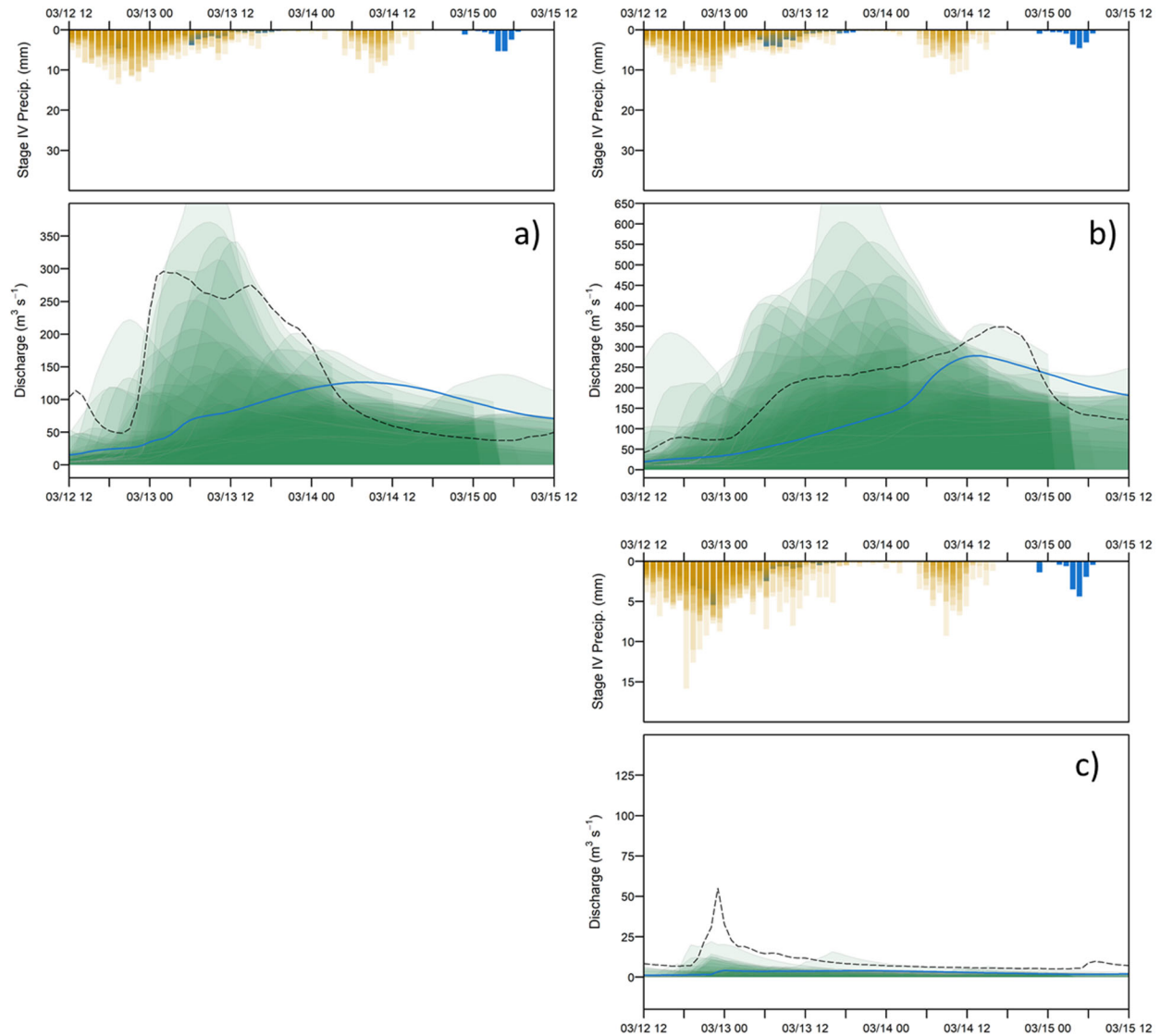


Figure 6.14: a) Ensemble forecast streamflow performance at FLW LWC T initialized at 12 March 2021, 00Z. The top plot is the hyetograph of hourly Stage IV precipitation (blue) and individual ensemble precipitation (gold) averaged across the area of the Roubidoux Creek catchment upstream of LWC T. The bottom plot shows the observed streamflow (black dashed line), the simulation forced by Stage IV precipitation (blue line) and each green shading represents a streamflow time series corresponding to one of the ensemble forecast members. (b)-(c) as in (a), but for Eastgate on the Big Piney River and LWC FLWU on Turnbull Hollow, respectively.

Figure 6.15 shows the Stage IV and forecast precipitation hyetographs and observed, simulated, and forecast hydrographs initialized at 30 June 2021, 12Z for Little Fountain Creek at Fort Carson. The forecast window corresponds to a minor flow event. The precipitation forecast ensemble shows precipitation occurring throughout the forecast window compared to the observed

precipitation, which occurred in distinct events. The resulting streamflow ensemble agreement (darker green shade) outperforms the Stage IV simulation, which overestimates the flow.

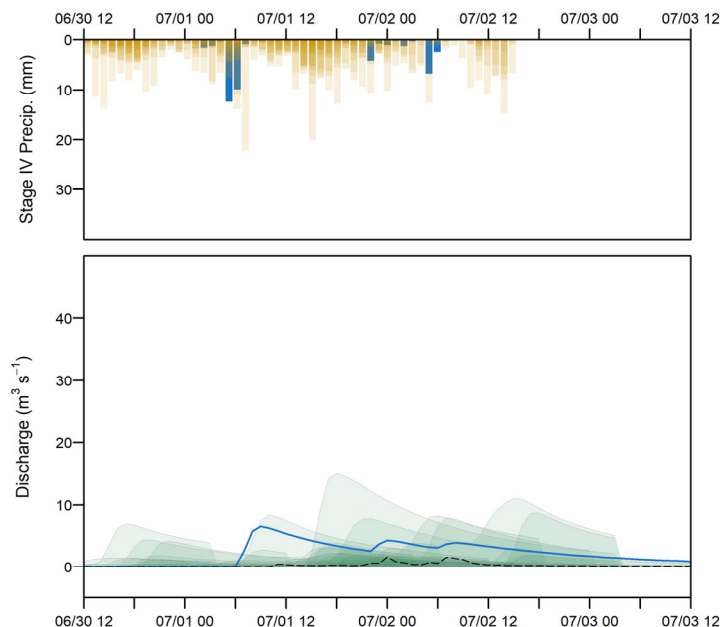


Figure 6.15: a) Ensemble forecast streamflow performance at Fort Carson LWC TTAlpha initialized at 30 June 2021 at 12Z. The top plot is the hyetograph of hourly Stage IV precipitation (blue) and individual ensemble precipitation (gold) averaged across the area of the Little Fountain Creek catchment upstream of LWC TTAlpha. The bottom plot shows the observed streamflow (black dashed line), the simulation forced by Stage IV precipitation (blue line) and each green shading represents a streamflow time series corresponding to one of the ensemble forecast members.

Skill score statistics are weighted heavily by the accurate prediction of the null events. A more in-depth event-based review is required to better understand the ensemble performance during active periods. Nonetheless, the performance of the forecast streamflow ensemble reflects the shortcomings of both the forecast precipitation ensemble and the calibrated HEC-HMS model concerning the timing and quantity of forecast flow. LWC infrastructure can become hazardous during relatively frequent (small) flood events, meaning the ensemble hazard forecasts generated from the ensemble streamflow forecasts may outperform the latter.

6.3 Downscaling of ensemble members with good computational efficiency

Section Summary:

1. The inclusion of WRF in the ensemble adds value by increasing the forecast horizon to 72 hours without significantly changing the performance of the ensemble.

Table 6.3 provides the resultant Brier Scores for each installation hazard site and summarizes 2-day or 3-day ensemble performances (i.e., the time span covered by all ensemble configurations considered). The resultant Brier Scores for all time periods suggest that all of the configurations

considered are rather skillful. With the exception of one of the Roubidoux sites at Fort Leonard Wood (FLW03, FLWT_Roub), all of the other scores were below 0.5, and most were less than 0.1. We note, however, that flood hazards are not a common occurrence, so the Brier Scores may be inflated by the accurate prediction of the null events. As such, these results should be viewed as preliminary guidance for ensemble design rather than a definitive measure of model performance.

The performance metrics from this assessment suggest that the baseline configuration forced with readily-available, convective-resolving atmospheric datasets may be sufficient for 48-hour flood risk forecast needs. However, the ensembles, which include both the convective-resolving and downscaled members (eg, Ens 11), performed marginally better at most locations than convective-resolving ensemble only (Ens 1). The downscaled members only ensemble (Ens 6) performed marginally worse than Ens 11, but included all 72 forecast hours compared to 48 hrs. The last 24 hrs of the 72 hr forecast are the most uncertain and the most likely to diverge from reality, nonetheless, the BS is only slightly worse when including these hours, indicating that the downscaled members add utility to by increasing the forecast horizon to 72 hrs without significantly changing the performance of the ensemble. The Brier Scores at all sites remained relatively consistent for the ensemble configurations that extended out to 72 hours, suggesting that downscaling may be worth the effort if installation managers and decision makers find value in the extra day of lead time.

Table 6.3: Installation site hazard 2-day or 3-day ensemble prediction Brier scores, including the 03-48 forecast hour period for the 00Z and 12Z cycle NAM-CONUS-nest, HiResW, and WRF-GFS members and the 00-72 forecast hour period for the 03Z and 15Z cycle WRF-

SREF members. We used the Brier scores for Ens1 (grey), which is comprised of members that do not require downscaling via the WRF model, as a baseline for comparison. Ensemble configurations that matched the performance of Ens1 are highlighted in yellow, and configurations that improved upon the Ens1 score are highlighted in green. The lowest BS value for each site is bolded, if applicable.

	Convective resolving members only (48hr)	Downscaled members only (72hr)					Downscaled and convective-resolving members (48hr)				
SiteID	Ens 1	Ens 2	Ens 3	Ens 4	Ens 5	Ens 6	Ens 7	Ens 8	Ens 9	Ens 10	Ens 11
FC01	0	0	0	0	0	0	0	0	0	0	0
FC02	0.0010	0.0001	0	0	0	0	0.0001	0	0	0	0
FC03	0.0001	0	0	0	0	0	0	0	0	0	0
FC04	0.0001	0	0	0	0	0	0	0	0	0	0
FC05	0	0	0	0	0	0	0	0	0	0	0
FH01	0.0059	0.0030	0.0025	0.0024	0.0023	0.0022	0.0024	0.0023	0.0022	0.0022	0.0021
FH02	0	0	0	0	0	0	0	0	0	0	0
FH03	0	0	0	0	0	0	0	0	0	0	0
FH04	0.0150	0.0122	0.0116	0.0112	0.0112	0.0110	0.0111	0.0109	0.0108	0.0108	0.0107
FH05	0.0159	0.0108	0.0101	0.0097	0.0097	0.0095	0.0095	0.0093	0.0092	0.0092	0.0091
FH06	0.0353	0.0324	0.0315	0.0314	0.0314	0.0311	0.0305	0.0303	0.0304	0.0305	0.0304
FH07	0.0169	0.0063	0.0054	0.0049	0.0048	0.0046	0.0047	0.0045	0.0042	0.0043	0.0042
FH08	0.0155	0.0057	0.0048	0.0043	0.0043	0.0041	0.0042	0.0040	0.0038	0.0038	0.0037
FH09	0.0110	0.0044	0.0035	0.0031	0.0030	0.0029	0.0032	0.0029	0.0028	0.0027	0.0027
FH10	0.0124	0.0047	0.0038	0.0034	0.0033	0.0032	0.0034	0.0031	0.0030	0.0030	0.0029
FH11	0.0142	0.0098	0.0090	0.0087	0.0086	0.0084	0.0085	0.0083	0.0082	0.0082	0.0081
FH12	0.0209	0.0153	0.0145	0.0139	0.0139	0.0137	0.0133	0.0133	0.0131	0.0132	0.0131
FH13	0.0160	0.0121	0.0113	0.0109	0.0108	0.0107	0.0105	0.0104	0.0102	0.0103	0.0102
FH14	0.0209	0.0153	0.0145	0.0139	0.0139	0.0137	0.0133	0.0133	0.0131	0.0132	0.0131
FH15	0.0477	0.0408	0.0404	0.0398	0.0401	0.0399	0.0356	0.0369	0.0372	0.0378	0.0381
FH16	0.2143	0.2399	0.2435	0.2434	0.2440	0.2451	0.2346	0.2394	0.2406	0.2421	0.2439
FH17	0.0602	0.0531	0.0530	0.0525	0.0528	0.0526	0.0469	0.0487	0.0492	0.0501	0.0506
FH18	0.0468	0.0342	0.0326	0.0322	0.0323	0.0320	0.0287	0.0294	0.0298	0.0304	0.0306
FH19	0.0485	0.0387	0.0376	0.0371	0.0373	0.0370	0.0331	0.0342	0.0346	0.0353	0.0356
FH20	0.0153	0.0141	0.0137	0.0136	0.0135	0.0135	0.0134	0.0133	0.0132	0.0132	0.0132
FLW01	0.0001	0.0003	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001
FLW02	0.0971	0.0477	0.0508	0.0536	0.0547	0.0552	0.0418	0.0435	0.0464	0.0481	0.0495

FLW03	0.6593	0.7901	0.8369	0.8433	0.8500	0.8562	0.7899	0.8198	0.8272	0.8356	0.8437
FLW04	0.0898	0.1001	0.1000	0.0997	0.0998	0.0998	0.0983	0.0983	0.0981	0.0983	0.0983
FLW05	0.0056	0.0063	0.0057	0.0058	0.0057	0.0056	0.0056	0.0054	0.0055	0.0055	0.0054
FLW06	0.2236	0.2332	0.2769	0.2935	0.3008	0.3047	0.2058	0.2376	0.2562	0.2678	0.2763
FLW07	0.3915	0.6452	0.7339	0.7527	0.7651	0.7758	0.6359	0.6949	0.7160	0.7327	0.7482
FLW08	0.1685	0.2554	0.2792	0.2857	0.2893	0.2915	0.2435	0.2614	0.2689	0.2743	0.2783
FLW09	0.0001	0.0003	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001
FLW10	0.0007	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001
FLW11	0.0001	0.0005	0.0002	0.0002	0.0002	0.0001	0.0002	0.0001	0.0001	0.0001	0.0001
FLW12	0	0	0	0	0	0	0	0	0	0	0
FLW13	0	0	0	0	0	0	0	0	0	0	0

Figure 6.16 shows the difference in hypothesis testing analysis between the full member ensemble with and without the downscaled members. The plots show the difference (with and without WRF downscaled members) in the percentage of timesteps of flood (True/True), simulated flood and observed no flood (False Positive i.e., Type I error), simulated no flood and observed flood (False negative i.e., Type II error), and simulated no flood and observed no flood (False/False). The results show no significant change with or without the WRF members, though the occurrence of F.F. decreases and the occurrence of Type I error increases, likely due to erroneous precipitation added by the WRF members. Both the BS and hypothesis testing indicates that inclusion of WRF in the ensemble adds value by increasing the forecast horizon to 72 hours without significantly changing the performance of the ensemble.

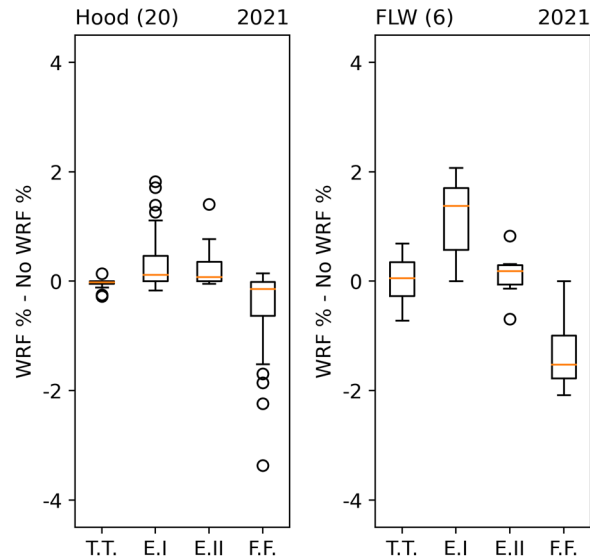


Figure 6.16: Difference in hypothesis testing analysis using ensembles with/without the WRF downscaling simulations for 20 LWCs at Fort Hood (left) and 6 LWCs at Fort Leonard Wood (right). Individual boxes are percent of 2021 time period simulated and observed flooding (T.T.), false positives (E.I), false negatives (E.II), simulated and observed quiescent (F.F.).

6.4 Forecasted risk agreement with historic conditions

Section Summary:

1. The infrastructure type, and subsequent hazard parameter used, is a key consideration when evaluating the model's ability to forecast historical conditions.
2. From a macro perspective, aggregating from a 1-hourly time step to a coarser temporal resolution does not show significant improvements in forecast errors (increased True/True outcomes and decreased Type I & II errors)
3. Guidelines for hydraulic stability thresholds are physically based and well-documented, but our findings support that these thresholds or ensemble agreement thresholds can be adjusted to site conditions and could be used as a calibration parameter to greatly improve risk forecast skill.

A statistical analysis with hypothesis testing evaluates HydroRF's ability to successfully forecast hazardous conditions at individual crossings. By using hypothesis testing, the frequencies of HydroRF's successes and errors can be quantified. Minimizing false negatives, i.e., when observations show conditions warranting a crossing closure but forecasts show safe conditions, is the most important with flood forecasting. Minimizing false positives is key for bolstering confidence in the model output, but false negatives can lead to the loss of life and property.

The analysis was conducted over the 2021 period coinciding with the WRF runs (March 09 - July 15). For every timestep and at every crossing, the observational flow data were used to compute if that timestep was hazardous or not. For the same time steps across all the hourly initializations, the same determination of hazardous conditions was computed for every flow ensemble member.

The comparison of forecasts to observations effectively produces four outcomes: both showing flood (True/True), simulated flood and observed no flood (False Positive i.e., Type I error), simulated no flood and observed flood (False negative i.e., Type II error), and simulated no flood and observed no flood (False/False). The frequencies of these four bins was calculated for each crossing at each installation. Boxplots of similar infrastructure types were then made to depict the four outcome percentages of the total timesteps. Fort Carson is not included due to the very small sample size.

At FLW, two different safety parameters were used. The hydraulic stability threshold (HST) is more appropriate at the LWCs, while the water surface elevation (WSE) representing inundation is more appropriate for range infrastructure. Inundation of range infrastructure (located in the flood plains) is rare compared to the frequency of the LWC driving surfaces being overtopped (Figure 6.17). During 2021 at the HST crossings, the median percent of timesteps that were both simulated and observed quiescent is 52%. At range infrastructure WSE crossing, this value is 99%. Forecasting floods at the LWCs is a more frequent problem which directly translates to a greater risk, therefore results are confined to only the LWCs herein. No range infrastructure was inundated at FLW during 2020 or 2021.

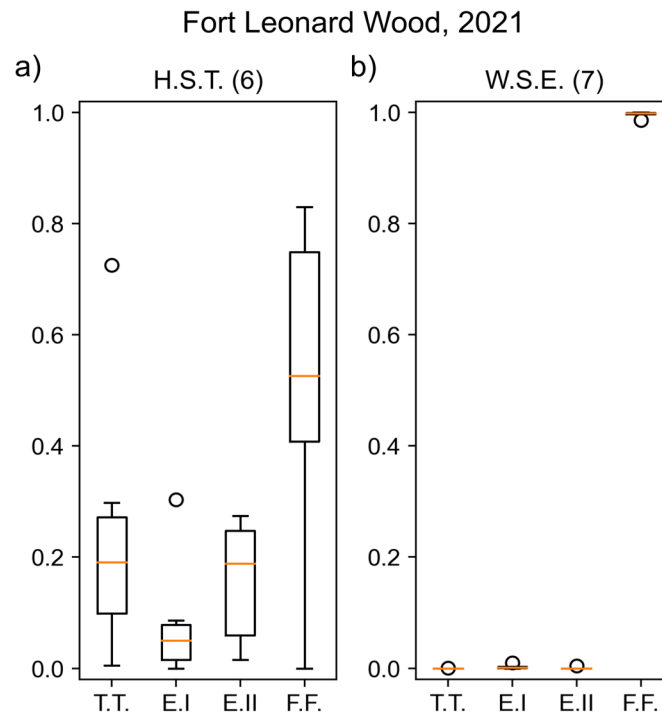


Figure 6.17: Hypothesis testing at Fort Leonard Wood for March-July 2021 with an ensemble of NAMnest, HRRR, HRRRE, HiResW, and WRF members. (a) Boxplot of percentages of true/true, type I error, type II error, and false/false timesteps at the 6 hydraulic stability threshold LWCs at FLW. (b) The same as (a) for the 7 water surface elevation range infrastructures.

The performance of HydroRF at FLW LWCs is mixed (Figure 6.18a). There is considerable spread across the 6 LWCs, but when aggregated together the average percentage of timesteps correctly simulated as hazardous flooding is 25%. That combined with the mean correct quiescent

percentage (52%) results in an average of HydroRF correctly forecasting 77% of the 2021 timesteps. The remaining 23% of the mean behavior is attributed to 8% false positives and 15% false negatives. Having more “true” hits than errors supports HydroRF’s value, but ideally the false negatives would be less than the false positives. From an operational perspective, this is where the calibration of the HST values presents an opportunity to compensate for some of the errors that are ingested into the model chain, such as precipitation uncertainty and capturing hydrologic processes. The hypothesis testing was conducted using literature values of 0.4-0.6, but the sensitivity of the hypothesis testing to these values will be quantified in subsequent sections.

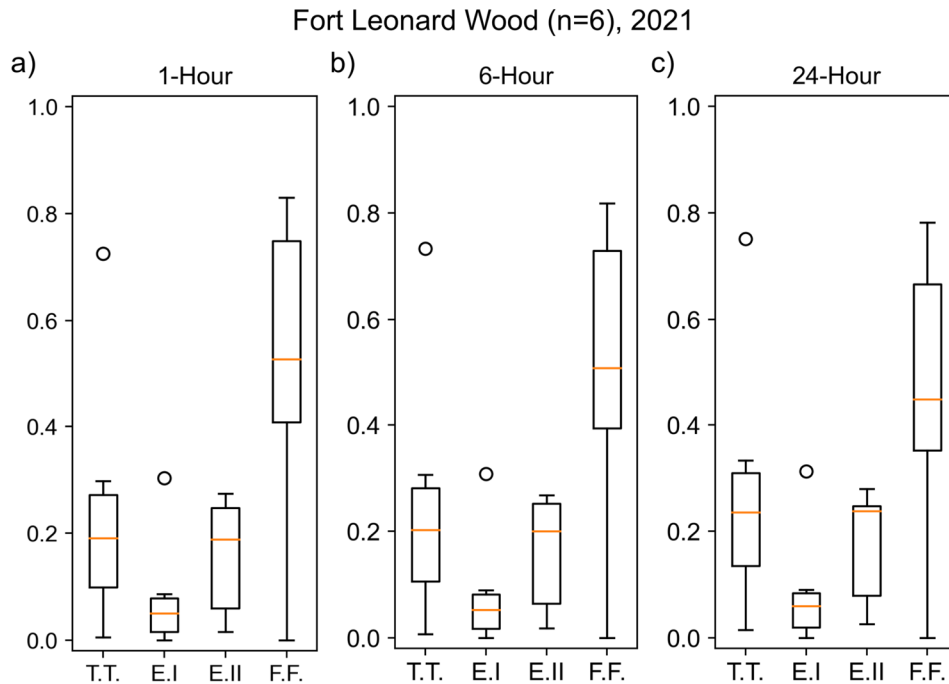


Figure 6.18: Temporal aggregation with duration maximums for hypothesis testing at Fort Leonard Wood for March-July 2021. (a) control with original hourly data (b) 6-hourly aggregated maximums (c) daily aggregated maximums.

The reporting interval (e.g., hourly, 6-hourly, daily) of HydroRF is an aspect that can vary and be set based on stakeholder preference. The hypothesis testing analysis was repeated with time series aggregated to match these possible reporting intervals to examine aggregation’s effect on the risk reporting performance. The aggregation was done by taking the maximum streamflow value for a given timeframe, similar to how peak flows are determined from annual data. At both FLW (Figure 6.18) and Fort Hood (Figure 6.19), the aggregation results in a reduction of the simulated/observed quiescent percentage. This result is intuitive because the aggregation by maximums should result in more emphasis on any hydrologic activity present that is not baseflow conditions. At FLW (Figure 6.18), the median of the false/false bin decreases from 52% at hourly to 45% at daily temporal resolution. The median of the true/true bin increases from 19% to 24%, but the false negative bin also increases from 19% to 24%. Similar results occur with the Fort Hood simulations, but are more dampened due to the limited number of hydrologic events that occurred at Fort Hood in 2021. Due to the data limitations, these results support that a continuum-based analysis such as this should be complemented by an event-based analysis to capture a more holistic evaluation of

model performance. The aggregation does support that coarsening the temporal resolution of HydroRF does not result in significant improvements in model performance though.

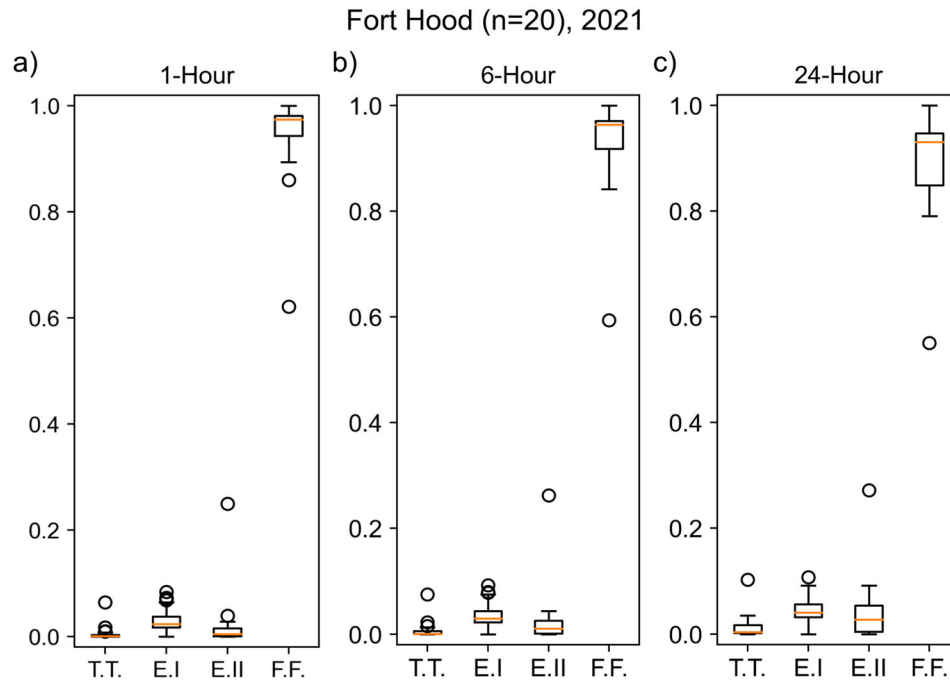


Figure 6.19: Same as Figure 6.18, but for Fort Hood

The sensitivity of hypothesis testing results was examined using two complementary analyses. The first analysis used the streamflow observations to determine all of the timesteps that were actually observed hazardous (denoted as observed HST greater than HST from literature). For all of those timesteps, the simulated HST (i.e., the multiplication of simulated flow depth and velocity) was computed for all of the ensemble members. The analysis aims to identify if there are any biases in the HST values on a site-by-site basis, which can be used to calibrate the HST values individually. Figure 6.20 shows the variability of HST values during observed hazardous conditions for the 5 LWCs on Roubidoux Creek at FLW. It is evident that the literature values are appropriate for the LWCs FLW 8 and Springfield Road where the median computed HST values are nearly 1.0. The median for FLW T and Stillwater Road are both below $0.4 \text{ m}^2 \text{ s}$ which suggests that a value smaller than the literature recommended would be more appropriate with regards to maximizing hazard forecast performance if treating HST as a calibration value.

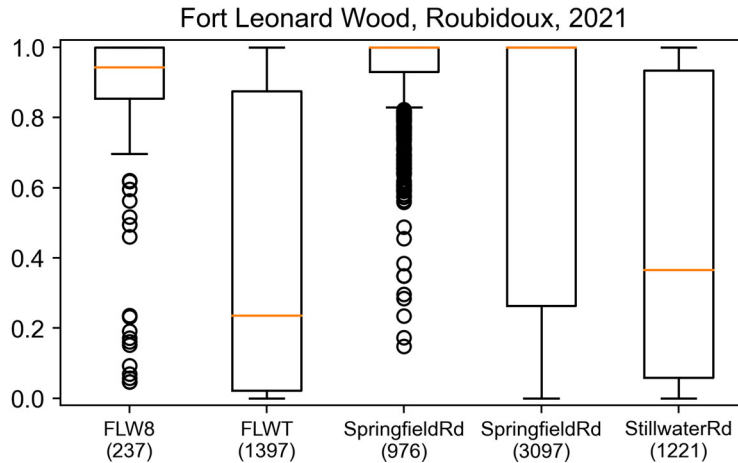


Figure 6.20: Boxplot of simulated HST values ($\text{m}^2 \text{s}$) during 2021 hazardous conditions based on observations for the 5 LWCs located on Roubidoux Creek

At Fort Hood, the computed HSTs are consistently below the literature recommended value of $0.4 \text{ m}^2 \text{s}$ (Figure 6.21). Table Rock Creek is constrained by limited observations, but it is clear that during observed hazardous conditions the entire ensemble is underestimating the flood hazard. The underestimation can be attributed to the small area of the Table Rock Creek watershed where the precipitation forecasts are not skillful at correctly placing precipitation and the HEC-HMS model is not skillful at capturing the hydrologic processes. Cowhouse Creek, which is a significantly larger watershed, shows a more consistent pattern of median computed HST values being near the literature recommended of $0.4 \text{ m}^2 \text{s}$. The analysis as a whole supports that a site-by-site calibration of HST values could be useful to maximize accuracy of the forecasted hazard assessment, however, as HST values are based on hazardous conditions measured from certain vehicles or pedestrians, an alternative calibration parameter is also investigated.

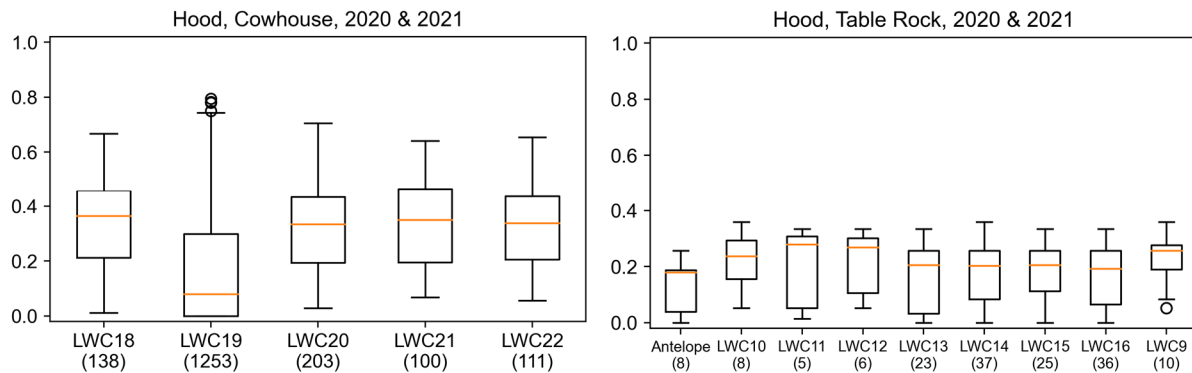


Figure 6.21: Same as Figure 6.20, but for Cowhouse Creek at Fort Hood (left) and Table Rock Creek at Fort Hood (right).

The second analysis investigates the probabilistic component of the risk communication which is determined by the percentage of ensemble members that are showing hazardous conditions for a given timestep. This is separate from the HST threshold which is used to define what a hazardous condition is. The baseline confidence value was 50% of the ensemble members showing agreement.

Figure 6.22 is an example of two LWCs at FLW where one site is not appropriate (left plot) to use the baseline value and one is appropriate (right plot). At FLW-T, a more appropriate confidence value to warrant mitigating action would be around 30% where true/true counts are higher than both type I and II errors. At FLW-8 at 50%, false negatives and positives are being minimized, while the true/true count is still greater than both. The correct confidence threshold is not as explicit at Fort Hood (Figure 6.23) because there is never a percentage where the true/true count is greater than both false positives and false negatives. In an effort to minimize false negatives, the confidence threshold should be lower around 20% with the caveat that false positives increase greatly. Greater false positives can result in stakeholders having less faith in the model itself making it more difficult to convince stakeholders to use HydroRF to inform decision making. Nonetheless, site-specific confidence threshold could be used as calibration parameter to improve hazard forecast performance.

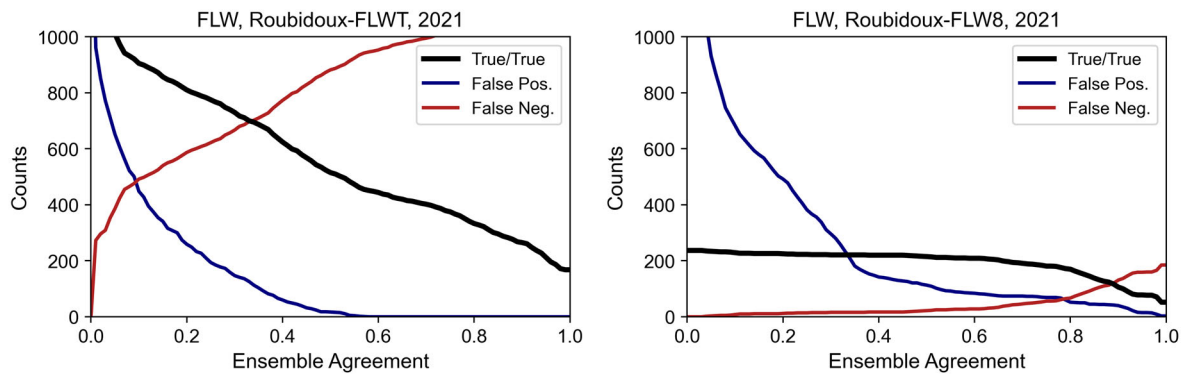


Figure 6.22: Counts of true/true, false positives, and false negatives for 2 locations at Fort Leonard Wood for varying ensemble agreement which is the percentage of the ensemble members that exceeds the HST threshold.

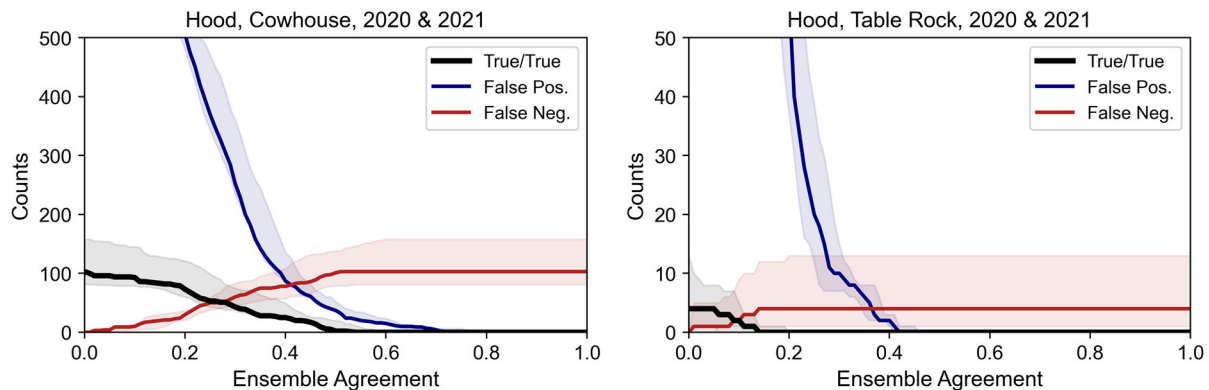


Figure 6.23: Same as Figure 6.22, but for Fort Hood where the median and interquartile range is shaded matching the sites included in Figure 6.21.

6.5 Risk report ease of use

Section Summary:

1. Respondents indicated that a forecasting hazard approach would be a useful complement to existing operational or passive hazard reports and would be of operational value.
2. Respondents indicated that communication strategies need to be as simple as possible, conveying required information without needing additional explanations.
3. Operationally, hazard forecasts should be updated every 6-12 hrs and be made available in the morning (0600 local time) and in the afternoon (1400-1600 local time).

The goal of this project is to demonstrate and validate a tool for Range Control, DPW, DES, Emergency Management, Operations, or other installation decision makers that uses near-term forecasts to limit risk to soldiers and installation training/testing missions by enabling the decision makers to take mitigation actions prior to the start of potentially unsafe or damaging flood conditions. Risk reporting is only effective if the decision makers can easily understand the report, appreciate the probable risks, and act accordingly to mitigate the risks. As a result, three preliminary risk assessment report communication strategies were presented to installation decision makers and their feedback was recorded via a fillable pdf survey form (Appendix B). The presentation occurred via Webex on 18 July 2022 . The webinar was facilitated by the Sustainable Range Program, G7, Live Training Branch, US Army Installation Management Command. Seventeen individuals from 12 US Army Installations participated, including ITAM coordinators, training range managers, Directorate of Plans, Training, Mobilization, and Security (DPTMS) chiefs, and range safety managers. A total of eight completed surveys were returned and responses from four other individuals were received by email or over the phone. A two-way conversation approach between developers and potential decision makers was pursued in order to develop refined forecast risk communication strategies (Figure 6.24). The goals of the risk communication strategy figures are to be easily understood, accurately demonstrate the probable forecast risk, and be actionable (enable decision makers to act accordingly to mitigate the risks).

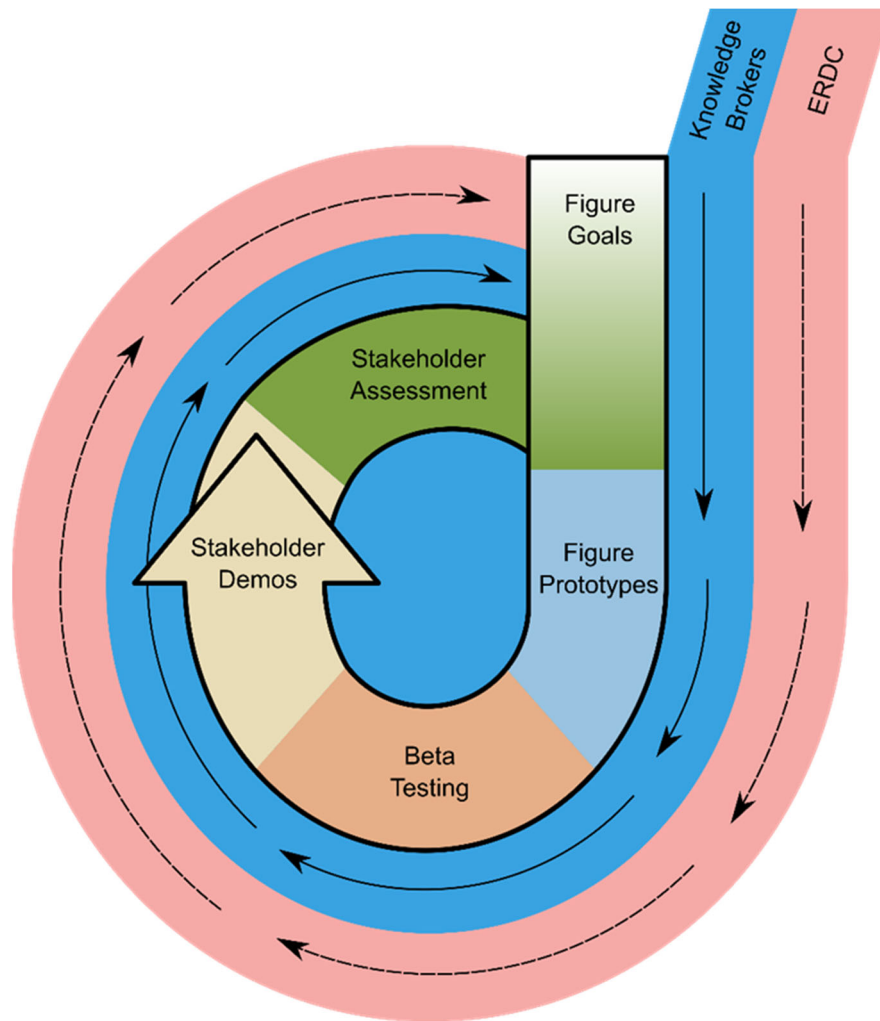


Figure 6.24: Two-way communication approach between risk communication strategy developers (ERDC) and potential installation decision makers (Knowledge Brokers) (adapted from Pennell et al., 2013).

Three prototype figures were created to provide decision makers with a range of detail and information with regards to the timing, confidence, and cause of the forecasted risk. The three prototype communication strategy figures presented to the knowledge brokers are described and introduced in Section 2.1.3. The fillable pdf survey form asked for responses to questions about general communication strategies, the three prototype communication strategies, and general questions related to operational considerations.

All of the respondents indicated that a forecasting hazard approach would be a useful complement to existing operational or passive hazard reports and would be of operational value for Range Ops, Range Safety, Training Operations, unit commanders, and/or Installation Commanders. Respondents were neutral with regards to the value of GO/NO GO (green/red) hazards reports compared to the GO/SLOW GO/NO GO (green/amber/red) hazard reports.

All respondents agreed that the tabular figure (Figure 5.19) is easy to read and understand and provides the necessary forecast information for operational application. Responses were mixed (about half agreed and half disagreed) about whether the tabular figure justified closing LWC 21 during the forecast window and also mixed about if the figure provided enough precise information about the timing and magnitude of the forecasted hazard conditions. A number of additional comments were received related to the tabular figure, the following comment is representative of the general tenor, *“Keeping the data in simplistic form allows Commanders and their staff to get a quick read on what may or may not be a hazard to their training plans. The more in depth the information the more frustrated commanders will be.”*

About half of the respondents indicated that they understood the heatmap figure, the other half strongly indicated that they didn’t understand it, and all agreed it is too complicated (Figure 5.20). However, most agreed that there is value in conveying in more detail how the magnitude and confidence of the hazard forecast varies in time. Responses mostly disagreed that the heatmap figure justified closing LWC 21 during the forecast window. The following comment is representative of the general tenor received regarding the heatmap prototype, *“The information presented is good data; however, in the context of Range Operations (Fire Desk Personnel, Range Inspectors, etc.) or Commanders it needs to be as simple as possible.”*

Most of the respondents indicated that they didn’t understand the rating curve figure and all agreed it is too complicated and that the lack of timing information makes the figure confusing (Figure 5.21). Nonetheless, most respondents agreed that a figure conveying the forecasted water velocity and depth would have value. Responses mostly disagreed that the rating curve figure justified closing LWC 21 during the forecast window but agreed it is too complicated. The following comment is representative of the general tenor received regarding the rating curve prototype, *“The information presented is good data; however, in the context of Range Operations end users (ie: Fire Desk Personnel, Range Inspectors, unit OIC/RSO's etc.) it needs to be as simple as possible.”*

The frequency and timing considerations of the operational forecasts were also considered by respondents. All agreed that new forecasts should be made available at least every 12 hours, however, there was at least one recommendation for each of the following frequencies: 12 hr, 6 hr, 3 hr, and hourly. Severely respondents suggested that the frequency of forecast reports change between quiet and active periods, such as 12 hr or 6 hr reporting intervals during quiet forecasts, increasing to 6 hr or 3 hr reporting intervals, respectively, during active forecast periods.

Most respondents advised that updated forecasts be made available by 0600 local time or between 1400-1600 local time, or both. The rationale for the 0600 forecast is that it provides personnel with the most recent information available before most training or testing exercises have begun for the day, allowing time for dicimination of both the hazard assessment and mitigation actions before soldiers are put at risk. The midafternoon forecast timing was recommended to *“give commanders/planners reaction time to develop ‘work-around’ solutions to potentially closed crossings.”*

During the webinar personnel from several installations shared the current safety operations with regards to LWCs and flood hazard. Fort Sill, OK closes a LWC once the water depth over the

driving surface is six inches or deeper. No forecast capability is currently used.. Fort Hood, TX has a network of stream gauges. When the 3rd AF Weather Squadron reports a forecast of at least 2 inches of rain in a 12 hr period near Fort Hood, installation Operations starts actively monitoring the gauge network remotely. Once one gauge reports a water depth corresponding to 1 foot over the driving surface of the adjacent LWC, all LWCs on the installation are closed. LWCs closing information is disseminated as quickly as communications allow.

Lastly, two respondents asked if a single crossing could be assigned multiple hazard assessments based on a range of anticipated traffic types. For example, could two hazard forecasts be created for the same crossing, one for pedestrians and one for heavy vehicles, if a LWC is commonly used for both troop marches and tactical vehicles wet gap training. Several other respondents suggested the color green be used instead of blue for the heatmap and rating curve figures. These comments were also considered by the ERDC-CERL team when updating the prototype figures.

Given the stakeholder assessments described above, efforts focused on reassessing figure goals and figure design for the heatmap prototype, given that most respondents agreed that there is value in conveying in more detail how the magnitude and confidence of the hazard forecast varies in time, while all agreed that the prototype heatmap figure is too complicated. As a result, the heatmap figure was resigned (Figure 6.25). The hourly HST values for each ensemble member were removed to make the figure cleaner. The background shading was aggregated from hourly to 6 hr timesteps while the intensity of the background red/blue shows model agreement (confidence) with darker colors indicating more members above/below the stability threshold (denoted by the yellow line). The stability threshold is the original value obtained from literature (0.6). The confidence color ramp was adjusted using an ensemble agreement threshold of 20% obtained from a review of Figure 6.23. The stability threshold corresponding to the USGS observations is in black here but would not be present on operational forecasts. The figure alterations are designed to make the heatmap figure easier to understand while still providing decision makers a more detailed visual of how the magnitude and confidence of the hazard forecast varies in time over the 72 hr forecast horizon.

The team assessed that the tabular prototype figure requires no further refinement at this time given the uniformly positive reviews and feedback. Nonetheless, an updated tabular figure is presented which mirrors the updates described above for the heatmap figure (Figure 6.26). The tabular figure was changed from 12 hr timesteps to 6 hr timesteps and the ensemble agreement threshold which denotes a risky forecast was changed from 50% to 20%. Future refinement of prototype figures will occur during tech transfer.

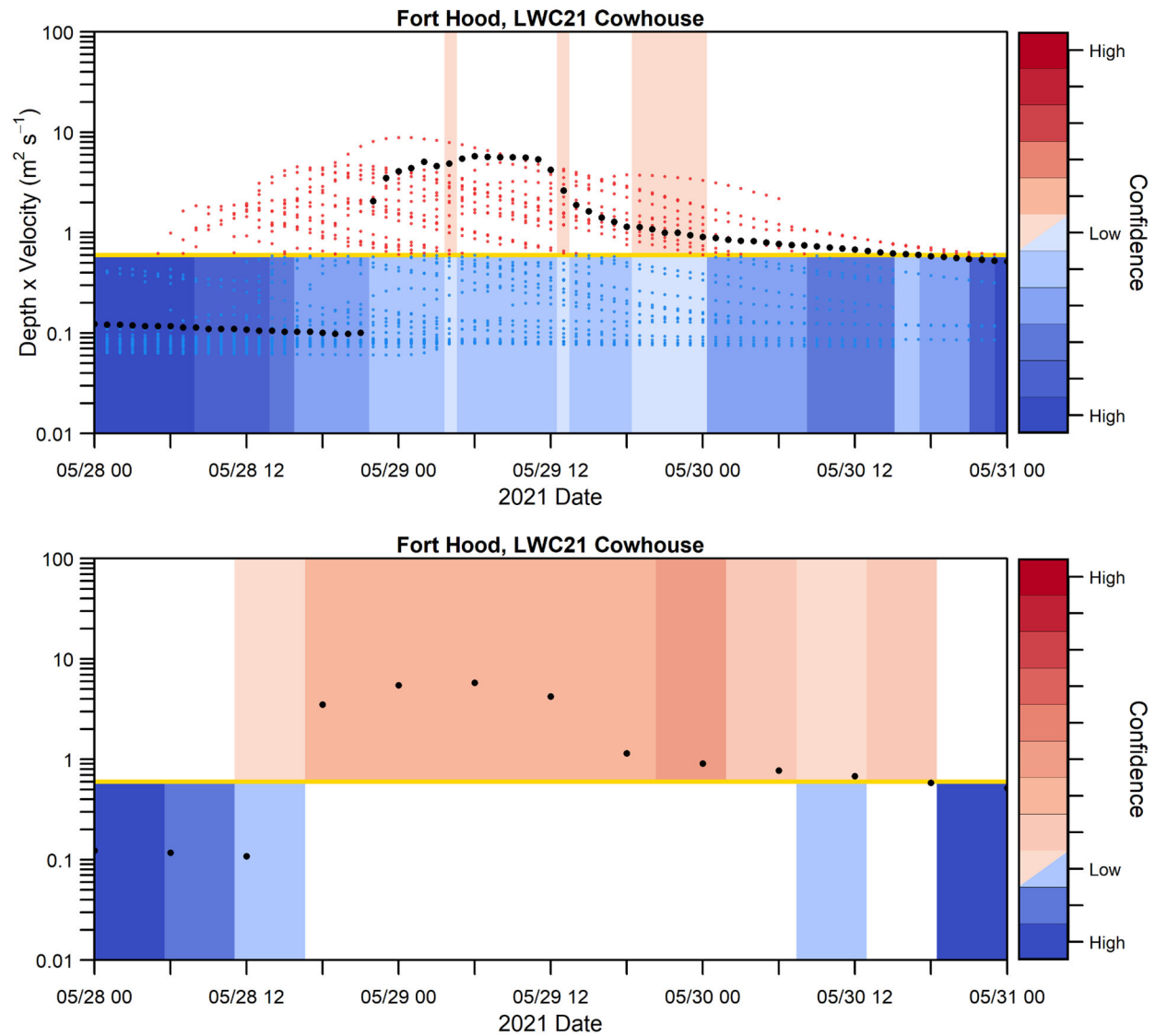


Figure 6.25: Example of the original heatmap figure (top) and modified (bottom) incorporating stakeholder feedback and ensemble confidence corrections taken from Section 6.3.

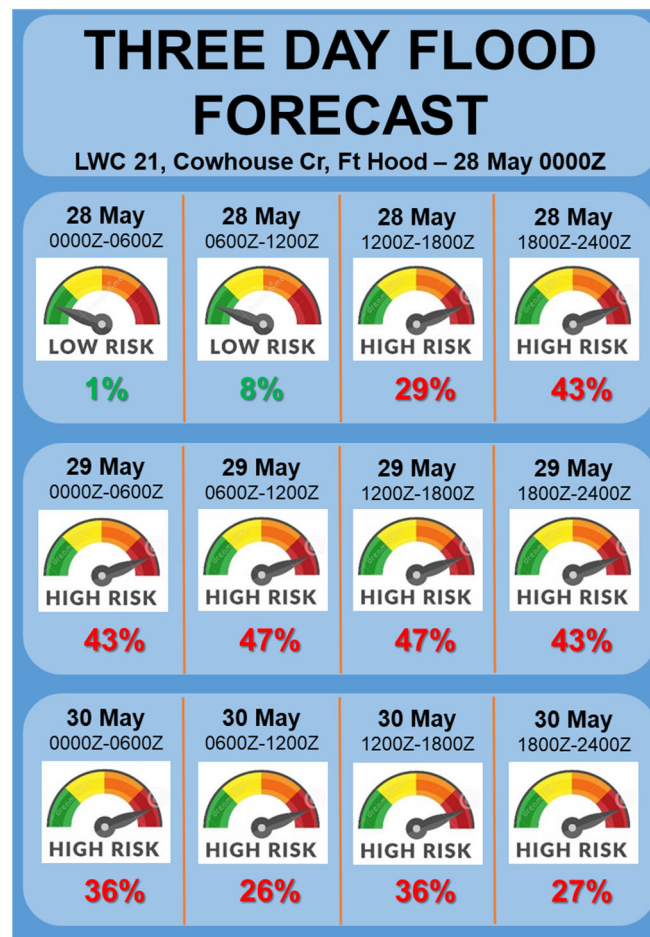
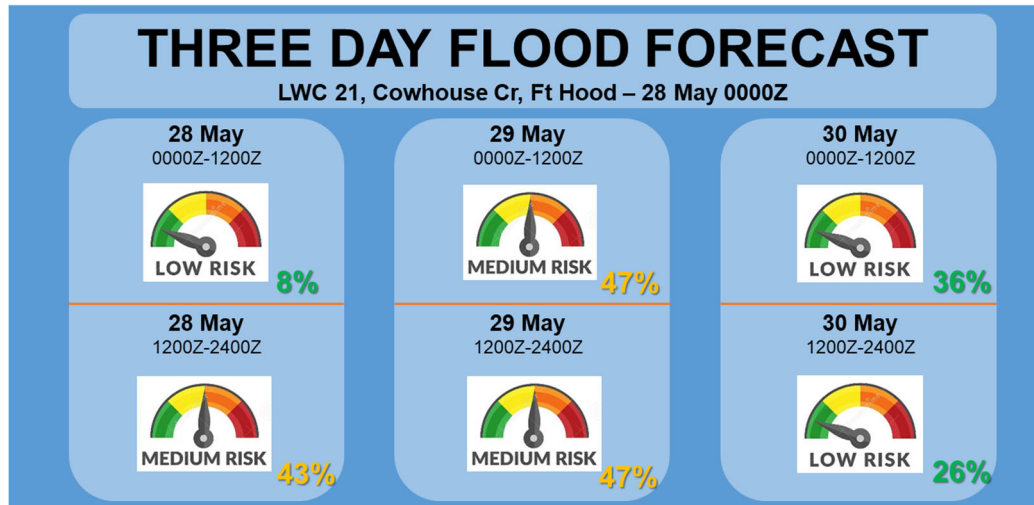


Figure 6.26: Example of the original tabular figure (top) and modified (bottom) incorporating stakeholder feedback and ensemble confidence corrections taken from Section 6.3.

Two methods are available if a single crossing needs to be assigned multiple hazard assessments based on a range of anticipated traffic types. If the HST value is not adjusted as part of a hazard calibration, HST values should follow the literature for different traffic types. If HST calibration is required for a site, the HST calibration can be completed for each anticipated traffic type by determining the observed hazard timesteps for various HST values from literature. Once each site is calibrated for each required traffic type, a hazard forecast for each can be added to the hazard report.

7.0 COST ASSESSMENT

The purpose of Hydro RF is to operationally predict flooding at critical infrastructure on military installations through the ingestion of probabilistic weather forecasts. Standing up Hydro RF at an installation will incur costs which together are the implementation costs per installation. The expected implementation costs must be known before an installation can plan to stand up Hydro RF while the life-cycle cost must be known to anticipate long-term operational cost. Implementation costs include fieldwork/survey, model development, and upload fees. Long-term costs include hosting fees and upkeep costs associated with new or modified infrastructure at an installation already utilizing Hydro RF.

Cost factors to be considered include fieldwork/surveying, model development and/or calibration for atmospheric, surface, and risk modules, maintenance or model upkeep costs, and upload and/or hosting fees. Each demonstration installation was unique in terms of the number and type of basins, river miles, and infrastructure locations surveyed and modeled (Table 7.1). Any potential future installation at which Hydro RF could be stood-up would be just as unique and potential implementation costs would reflect installation-specific characteristics. In order to better estimate implementation costs across diverse installations, costs are determined per river mile, per installation, per gauged basins, or per infrastructure site as appropriate. Flat mobilization fees are also estimated for tasks requiring travel and field work to account for related expenses such as per diem and airfare (Table 7.2).

Table 7.1: Cost Estimation Parameters for Each Demonstration Installation

Study Site	# of hazard sites	# of surveyed structures	Gauged calibration basins	River meters modeled	River miles modeled	# river miles per surveyed structures
Hood	20	26	1	51303	31.9	1.2
FLW	14	18	2	84969	52.8	2.9
Carson	12	16	1	30053	18.7	1.2
Total	46	60	4	166325	103.3	1.7

Table 7.2: Cost Model for a Hydro RF Implementation at an Installation

Cost Element	Data Tracked During the Demonstration
Survey/Field Work	Estimates made based on cost of survey(s) during demonstration (\$ per per river miles + mobilization)
Atmospheric Module	Estimates made based on cost of model(s) development during demonstration (\$ per installation)
Surface Module (HMS)	Estimates made based on cost of model(s) development and calibration during demonstration (\$ per gauged basin)
Surface Module (RAS)	Estimates made based on cost of model(s) development and calibration during demonstration (\$ per river mile)
Risk Module	Estimates made based on cost for different infrastructure types during demonstration (\$ per infrastructure site)
Maintenance or Model Upkeep	Estimates made based on labor requirements to survey and modify models or newly built infrastructure within modeled watersheds (\$ per infrastructure site + mobilization)
Upload and/or Hosting Fees	Determine based on estimated coding requirements (\$ per installation)

Survey and Field Work Tasks are estimated per river mile, assuming little to no base flow and approximately 1.2-1.5 river miles per surveyed infrastructure (ie, bridge, culvert, lwc, etc). The presence of deeper baseflow (averaging knee to chest deep) slows down surveying and increases cost. Surveying structures also requires significant time, so the number of infrastructure points would also affect final costs. Costs are estimated based on demonstration site survey requirements, including 1 long trip (2 surveyors) and 2 shorter trips (1 technician) to download HOBO data.

Cost of model modules were determined from the demonstration based on the number of hours required to build and calibrate each model. The code required to pre-process, run, and post-process the data for each module was created during the Dem/Val and associated development costs are excluded here. Costs related to the Atmospheric Module would include WRF model development and testing for a new installation and updating member archiving algorithms and are determined on a per installation basis. WRF updates include determining the bounds and geometry of new domains, making sure it telescopes over the desired area properly, and running tests on the domain to ensure there aren't any numeric errors relating to domain placement and interaction with other domains. Costs are estimated based on related Dem/Val tasks performed by qualified atmospheric scientists. If WRF set was not needed, the cost for this module would be reduced by 40%.

The costs associated with the Surface Module include building new HEC-HMS models and calibration. The cost driver for building and calibrating such models is the number of individual gauges/basins the model needs to be calibrated too. Therefore, cost estimates are provided on a per gauged basin basis. The cost drivers for building and calibrating HEC-RAS are the length of the modeled streams (river miles) and the number of structures modeled. An assumption of approximately 1.2-1.5 river miles per modeled structure (ie, bridge, culvert, lwc, etc) was used. Structures tend to be much more numerous in urban areas, so final cost estimates would be much greater when modeling such areas. The cost assessment here assumes rural river reaches. Costs are estimated based on demonstration site modeling requirements and tasks being completed by qualified hydrologists or hydraulic engineers.

Setting up the Risk Module is infrastructure site specific, requiring hydraulic and usage data. Selecting the model agreement threshold and adding each site to a final risk report is also included under these tasks, to be performed by a qualified hydraulic engineer.

Maintenance or model upkeep cost with regards to updating existing Hydro RF implementations are based on the labor required to update each module on a per infrastructure site basis. The cost for model maintenance includes mobilization due to the assumption that additional surveying would be required. Minor updates to the HEC-RAS model and the Risk Module are also included. This assessment assumes that the infrastructure site in question was structurally modified in some way or a new site is being added within the existing HEC-HMS basin. If the Hydro RF update required modifications or additions to the existing HEC-HMS models (including new calibration basins), update costs would be substantially higher.

Tech Transition costs are determined under the assumption that the 557th WW accepts the technology, in which case operational funding would be the responsibility of the AF. As such, operational costs and hosting fees are not included in this cost assessment. The technical transition of the technology to the AF is assessed here, including an analysis of what sort of software maturation process would be needed to make the code operationally-ready. Currently, every process required to run Hydro RF, except one, can be run on a HPC in a Linux environment. There is currently one Microsoft Windows executable file required to process gridded precipitation data into the file format required for HEC-HMS. For this Dem/Val individual processes were completed independently to allow for quality control. Most of the remaining work would focus on integrating the individual pieces into an operational workflow and testing. This work would be done by a qualified engineer and is estimated to require two months, plus potential computer resources to host the technology.

The Upload Cost is assessed under the assumption that the technology has already been successfully transferred to the AF. The costs associated with this task are limited to the process of integrating a new installation into an operational Hydro RF system. This work would be done by a qualified engineer and is estimated to require two weeks.

Table 7.3 provides the final cost assessment for each element by cost units.

Table 7.3: Element Costs by unit for Hydro RF Implementation at an Installation

Cost Element	Cost Units	Cost	Mobilization Cost
Survey/Field Work	\$ per river miles	\$900	\$5,000
Atmospheric Module	\$ per Installation	\$11,000*	
Surface Module (HMS)	\$ per gauged basins	\$24,000	
Surface Module (RAS)	\$ per river mile	\$480	
Risk Module	\$ per infrastructure site	\$375	
Maintenance or Model Upkeep	\$ per infrastructure site	\$3,500	\$3,000
Upload Cost	\$ per Installation	\$11,000	
Tech Transition	One time cost	\$70,000	

* \$6,600 if WRF setup not included.

In order to validate the cost assessment, it was applied to the three demonstration installations (assuming no prior work) (Table 7.4). These costs exclude any maintenance costs or the one-time technology transition expense. Based on the Dem/Val Hydro RF project and other recent projects at ERDC-CERL with tasks covering similar levels of work, these final cost estimates are reasonable.

Table 7.4: Estimated Hydro RF implementation cost for demonstration installations assuming no prior work has been completed.

Study Site	Setup Cost
Hood	\$103K
FLW	\$153K
Carson	\$81K

8.0 IMPLEMENTATION ISSUES

The implementation of Hydro RF is dependent on a number of unresolved factors, the most important being the computing resources required to run the tool operationally. For this study low-resolution forecast members (ie. SREF) were dynamically downscaled using WRF. The downscaling requires significant computing resources, necessitating use of a HPC. The downscaled ensemble members provided extensive utility, extending the forecast window out to 72 hrs, at least 24 hrs further than most of the high-resolution members, without significantly changing the performance of the ensemble.

The USAF Mesoscale Ensemble Prediction Suite (MEPS) provides 4 km high-resolution forecast members with forecast windows of 72-84 hrs depending on the location (Hacker et al., 2011). The project team staffed official data requests to the 557th Weather Wing several times during this project in order to gain access to the required MEPS output files, but all requests went unfulfilled. Ingesting the MEPS ensemble into Hydro RF would allow for a 72-84 hr probabilistic forecast without the need for dynamic downscaling, greatly reducing computational requirement.

Regardless of final computational requirements, housing Hydro RF at a single location and disseminating the forecasts to individual installations would streamline operational usage and maintenance. The 557th WW is a reasonable host given its existing operational requirements to provide both weather and hydrology forecasts to installations. The 557th WW also produces the operational MEPS ensemble. The project team have been in contact with LtCol Jeff Cunningham of Air Combat Command Weather Requirements (ACC/A5W), Jeff Cetola (ACC/A5W), Jack Floyd (ACC/A5W), Bonnie Brown (ACC/A5W), Corey Hummel (A3W), Kurt Schubeck (A3W), Krsti Arsenault (LIS-NASA), Jerry Wegiel (LIS-NASA), Augusto Getirana (LIS-NASA), and Scott Rheingrover (LIS-NASA). On-going discussions focus on how HydroRF fits into existing AF hydrology requirements (with AF requirements personnel) and preliminary discussions on tech transfer requirements (with LIS-NASA personnel).

Currently, every process required to run Hydro RF, except one, can be run on a HPC in a Linux environment. There is currently one Microsoft Windows executable file required to process gridded precipitation data into the file format required for HEC-HMS. Running Hydro RF on an operational HPC, even if not required for dynamically downscaling ensemble members, would help to ensure operational robustness and stability. A workaround for the executable file would need to be determined.

9.0 FUTURE WORK

The Hydro RF Dem/Val showed favorable results concerning accurate operational hazard forecasts with minimal false negative errors at many locations and infrastructure types, particularly within larger demonstration basins. Forecasted hazard accuracy is lower for periods of very intense rainfall over small basins. Both the precipitation forecasts and the calibrated HEC-HMS models contributed to underestimated probabilistic forecasted streamflow. However, once converted to probabilistic hazard forecasts, Type I/II errors occur during a minority of forecast timesteps. Future work is required, at this time, to determine when and why the majority of Type I/II errors occur and to implement and test proposed hazard calibration methods. Therefore, Hydro RF is not currently a plug-and-play ready tool. None-the-less, a wide variety of knowledge is ready to be transferred as part of a preliminary technology transfer. For example, the hazard assessment and communication methods investigated as part of this Dem/Val are preferable to the existing hazard determination methods currently used at certain installations.

Certain additional investigations are required to determine the utility of certain model modules and methods to create an operational tool. These investigations can be completed within a 12 month period. Tasks 1-3 require no new data. Task 4 requires an additional forecast ensemble to test the proposed bias corrections.

1. Hydro RF demonstrated that meeting with range and safety managers at installations helps foster a relationship between researchers and the end users of Hydro RF. The feedback meeting produced insightful feedback that was then used to tailor Hydro RF's flood reporting to best meet our clients needs. Continuing to foster the relationship with end users of Hydro RF to show them our continuing improvements, and to ask for their continued feedback would enable Hydro RF to continually evolve and prevent a disconnect between the research community and the user community. Additional work is required for future risk communication refinement.
2. A The largest source of uncertainty in Hydro RF's model chain originates from the operational weather forecasts' abilities to skillfully predict precipitation over the relatively small watersheds. The validation of Hydro RF has shown that the hydrology modeling introduces its own uncertainty, especially in the smallest tributary watersheds. Improving the skill of operational weather forecasts is an incredibly resource demanding endeavor. A study aimed at minimizing the uncertainty introduced by the HEC-HMS model by expanding the hydrologic simulations from a deterministic configuration (1 hydrology run for 1 precipitation forecast) to a stochastic configuration (several hydrologic runs for 1 precipitation forecast) could be easily accomplished with the work that has already been done. Incorporating perturbations to hydrologic initial conditions and first-order model parameters (which have both already been done as part of Hydro RF's calibration) with a HEC-HMS model forced by observations (e.g. Stage IV and MRMS) would produce an ensemble envelope that might perform better than a deterministic alternative.
3. One of the major takeaways from the validation of Hydro RF is that the inclusion of more data would improve the statistics that quantify Hydro RF's skillfulness across a continuum. That being said, there is potential value in assessing the current probabilistic

hazard forecasts on an event-by-event basis. Studying an event-by-event basis, and aggregating across those independent events, could produce several insights, such as when the majority of Type I/II errors occur, as well as why those errors are occurring based on the physical processes (both from an atmospheric and surface hydrology perspective). The timing of the errors could greatly impact the operational usefulness of Hydro RF. For example, do the majority of Type I errors occur during quiescent periods when numerous false positives could greatly undermine user confidence or adjacent to observed flood events when such errors would be less significant? From a physical processes perspective, the skill of Hydro RF could be evaluated for events that are identified as hydrologic responses to mesoscale convection systems and compared to those driven by front dynamics. Additionally, hydrologic events that are predominately infiltration excess runoff could be compared to those that are saturation excess runoff. Insights from an analysis such as this could be incorporated into the communication of risk to the end user by acknowledging that a certain physical process in the forecast is one that Hydro RF is historically proficient or inefficient at modeling.

4. Ultimately, Hydro RF demonstrated and validated that a model chain can be produced that utilizes probabilistic weather data to improve deterministic flood reporting. Thresholds, such as the HST and member agreement, are how Hydro RF collapses the probabilistic information of the ensemble into a deterministic decision for the end user. These thresholds should continue to evolve as the risk communication strategies evolve. Any threshold is arbitrary in nature, but that isn't to say that thresholds don't hold value. Additional work should focus on applying the HST and/or member confidence bias correction methods discussed in Sections 6.4 and 6.5. The tuning of these thresholds is one of the most cost effective ways of maximizing Hydro RF's value. For fairly little computational demand, these thresholds (and their sensitivities) can be further examined to find values that maximize Hydro RF's successes while minimizing the false positives and false negatives that are detrimental to end users confidence in the model chain and personnel safety, respectively. Additional forecast ensemble data would be required to test the effectiveness of the bias correction.

Closing the 1 hr forecast gap to allow for shorter term road closures is impractical at this time as all available forecast members have timesteps no shorter than 1 hr. The speed of information dissemination also prevents effective implementation of sub-hourly mitigation actions.