Article

Assessment of Relative Accuracy of TxRR and NWM in Simulating Streamflow for a Cluster of Watersheds along the south Texas Coast

Md Arifur Rahman[1], Yu Zhang[1], Mohammadvaghef Ghazvinian[1], Nelun Fernando[2], Caimee Schoenbaechler[2], and Yanjun Gan[1]

[1] Department of Civil Engineering, The University of Texas at Arlington, Arlington, Texas, USA.
[2] Texas Water Development Board, Austin, Texas, USA.
* Correspondence: mdarifur.rahman@mavs.uta.edu

ABSTRACT: Study Region -This study is conducted for a cluster of watersheds within the Matagorda Basin along the south Texas coast in the United States. **Study Focus** - Retrospective streamflow simulations of two hydrologic models of contrasting formulations and complexity, namely the TxRR, a simple analog model, and the National Water Model (NWM), a land surface-hydrologic model that explicitly accounts for surface energy balance in calculating water budget, and whose output. The comparison was motivated by a) the need for improving the modeling of runoff dynamics for watersheds along central Texas coast through the introduction of NWM to inform freshwater inflow monitoring and flood planning, and b) the need to better understand the potential ability of the NWM to address shortcomings of TxRR related to the latter's mechanistic deficiencies in a region known for hydroclimatic extremes. The study focuses on the temporal scale and climate dependence in the relative performance of the two models, the behaviors of models during extreme floods and droughts, and the relative efficacy of parameter transfer schemes. It further seeks to relate the performance differentials to model physics and calibration approaches. **New Hydrologic Insights** – NWM, with sophisticated representations of energe closure, is advantageous to TxRR in capturing low-frequency (interannual) variations in streamflow modulated by the surface energy balance, as evidenced by the superior correlation of its streamflow simulations at the interannual range. Yet, its overall performance trails behind TxRR at daily scale, and it tends to underpredict runoff volumes during two major floods. These features underscore the tendency of the model to suppress soil moisture at the onset of these events. On the other hand, TxRR outperforms in reproducing the volumes for the flooding events, but it overpredicts runoff during the extreme drought episode of 2011, likely an outcome of inadequate representation of the impacts of root zone soil moisture in regulating runoff. TxRR's parameteralization scheme proves more effective for adaptation across watersheds due to the presence of steep gradation in soil properties.

Keywords: Hydrologic model; ungauged catchments; coastal zones; water balance; parameter transferability

1. INTRODUCTION

Freshwater inflows to bays and estuaries regulate near-coast salinity, nutrient, and sediment dynamics, and thereby serve critical ecosystem functions (Russell *et al.*, 2006; Kim and Montagna, 2009). Observing terrestrial portion of freshwater inflow, however, is complicated by the fact that coastal streams are influenced by tidal currents and standard depth-discharge relations are often inapplicable. As a result, estimation of freshwater inflow often relies instead on hydrologic model simulations. Such models are typically calibrated over upstream stations free of tidal influence, and then the resulting parameter values are adapted to ungauged coastal watersheds.

doi:10.20944/preprints202210.0415.v1

The Texas Water Development Board (TWDB) operates a coastal modeling system that follows this paradigm. The hydrologic model in this system, the Texas Rainfall-Runoff (TxRR; Matsumoto, 1992), is a semi-empirical, semi-conceptual model that comprises a water balance module, a baseflow module and a Unit Hydrograph-based routing model. The water balance module of TxRR uses a modified version of the Soil Conservation Service (SCS; now NRCS) approach that allows for variation of the storage capacity in time (see similar works in Williams and LaSeur, 1976; Hawkins, 1978; Mishra and Singh, 2004). The strength of TxRR lies in its structural simplicity and parsimony in parameterization, which make it amenable to calibration. In addition, the model's use of the curve number allows one to estimate an initial set of parameters, or *a priori* parameters from physiographic features; it also offers a formal mechanism for transferring parameters across watersheds. Nonetheless, the lumped, conceptual nature of TxRR may have limited the ability of the model to capture key aspects of runoff dynamics. Potential limitations of TxRR include: 1) its lack of representation of interannual variations of meteorological conditions in determining losses; 2) its inability to account for subdaily and intra-watershed rainfall variability; and 3) its use of a somewhat crude mechanism to calculate "baseflow" without any differentiation of interflow and deeper groundwater discharge.

The past three decades saw the emergence of a number of gridded, physics-based, spatially distributed hydrologic models (Ogden and Julien, 1994; Cosgrove *et al.*, 2016). These models are able to account for spatio-temporal variability of rainfall, and offer sophisticated treatments of land surface processes. Among these, the National Water Model (NWM; Cosgrove *et al.*, 2016) has been operational at the National Weather Service since 2016. The model is an implementation of the WRF-Hydro (Gochis *et al.*, 2015), which represents a fusion between land surface and hydrologic modeling paradigms: it accounts for dynamic changes in land surface conditions in calculating surface water and energy fluxes; and incorporates representations of a wide range of surface and subsurface hydrologic processes.

As NWM is a relatively new model, to date only limited works have touched on its performance, and few of these focused on the skills of NWM vis-à-vis simpler, conceptual models (Nobre et al., 2016, Hansen *et al.*, 2019; Johnson *et al.*, 2019; Viterbo *et al.*, 2020a; 2020b). Loosely related, there is a body of literature on relative performance of physically-based, distributed and lumped, conceptual hydrologic models. These include some of the past model comparison experiments, notably the Model Parameter Estimation Experiment (MOPEX; Duan *et al.*, 2006), and the Distributed Model Intercomparison Projects (DMIP; Reed *et al.*, 2004; Smith *et al.*, 2012). The outcomes of these efforts illustrate that complex, physics-based models are not immune to under or misrepresentation of runoff processes, and their performance can be compromised by the uncertainties in parameterization and associated difficulties in model calibration. On the other hand, there has been evidence that the improved physical representation in WRF-Hydro leads to improved runoff-soil moisture coupling (Crow *et al.*, 2018).

Heretofore, much is unknown about potential advantages of energy-balance scheme in representing water balance for coastal/ near-coastal regions in regions that experience wide variations in climate regimes. The present study was conceived in recognition of this gap - it features close comparisons of NWM and the TxRR streamflow simulations for a cluster of watersheds situated in southern Texas coast which experienced some of the hydroclimatic extremes over the last few decades. The comparison will be done over a cascade of temporal windows (e.g., daily, monthly, and annual), and over extreme flood and drought episodes, thus allowing for an exploration of the merits of improved physical realism of NWM. It complements, and surpasses many extent ones by addressing the following research questions that have yet received their due attention, namely: 1) the differential skills of the two models in resolving flow conditions under extreme hydroclimatic conditions and the mechanistic underpinnings; and 2) the relative efficacy of NWM's parameter transfer scheme versus that of TxRR that is based entirely on physiographic similarity. The study no only serves the practical purpose of gauging the uncertainty associated with model-based runoff estimates for coastal, ungauged watersheds, it contributes to enriching our understanding of roles of land surface processes in shaping hydrologic response as well as potential tradeoffs between model structural complexity and parameter adaptability, both pivotal themes of contemporary hydrologic research (Duan *et al.*, 2006; Bardossy et al., 2007; Smith et al., 2012).

The remainder of this paper is organized into four sections. Following the introduction, the materials and methods are briefly introduced where the basic concepts of TxRR and NWM models are described in short. In the third section, result analysis and discussion of the model simulations are presented in detail. The last section summarizes the findings and concludes the work.

2. MATERIALS AND METHODS

2.1. Description of Model Physics

2.1.1. Texas Rainfall-Runoff (TxRR) Model

The Texas Rainfall-Runoff (TxRR) model is a lumped, semi-conceptual and semi-empirical model developed by TWDB. Presently, the agency is using the model to create daily and monthly runoff estimates for the ungauged watersheds that directly drain to the bays of Texas. These estimates are then combined with upstream inflow measured at USGS stations, diversion and return flow (discharge from wastewater treatment plants), and direct precipitation inputs to bays and estuaries to create total freshwater inflow estimates. Such estimates serve as input to a hydrodynamic model to estimate the current and salinity distributions in major estuary systems, which are used to inform water planning and the creation of environmental flow standards.

A simple schematic of the TxRR model is shown in Fig. 1. The model consists of three constituent modules, namely, a soil moisture accounting-based water balance scheme, a baseflow module, and a unit hydrograph-based routing model. The water balance scheme is used to calculate precipitation excess that is routed through the unit hydrograph, and the routed runoff is combined with baseflow to yield the total streamflow.

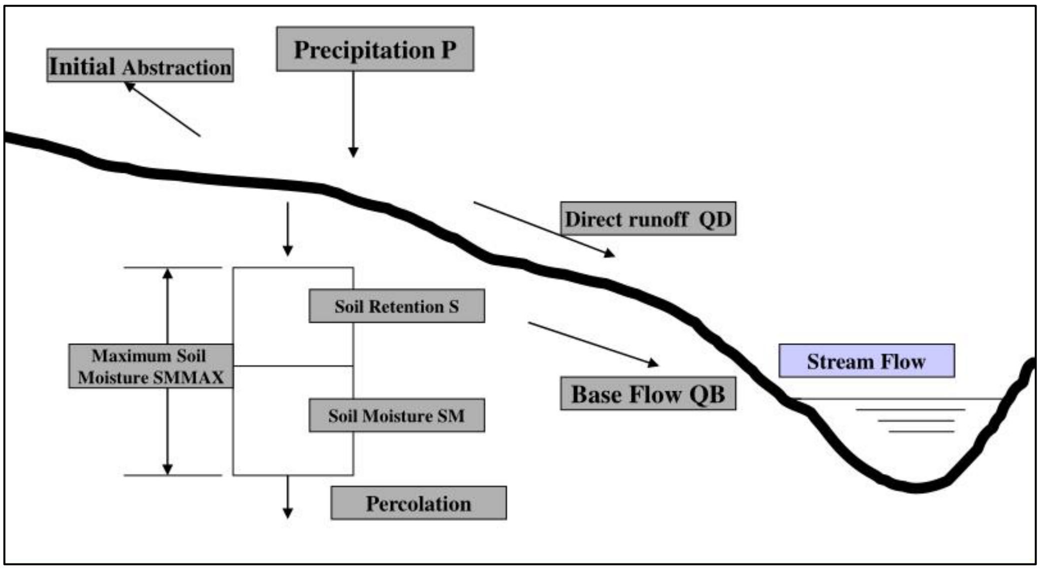


Figure 1: Schematic of TxRR model (Matsumoto *et al.*, 1994).

The baseflow and unit hydrograph models are similar to those featured in many of the contemporary modeling systems, including the USDA NRCS Technical Release No. 55 (TR-55; (Cronshey *et al.*, 1985), and US Army Corps of Engineers HEC-HMS (Scharffenberg, 2016). The water balance scheme is a tailored adoption of the SCS scheme. In the original SCS approach, for a given rainfall episode, the soil retention (S) is a static quantity that depends only on the curve number (CN). TxRR, by contrast, differentiates between soil retention and total soil water storage (SMMAX), an upper limit of retention. SMMAX is divided into soil retention (S), which represents the soil moisture deficit, and actual soil moisture (SM). The soil retention, S determines the initial abstraction and runoff using the SCS equation,

$$QD_i = \frac{(P_i - I_{ai})^2}{(P_i - I_{ai}) + S_i} \tag{1}$$

Where P_i is the total precipitation, and I_{ai} is the initial abstraction. Unlike the original SCS method, where initial abstraction is set to be 20% of S, in TxRR I_{ai} is related to soil retention S_i through a *tunable* abstraction parameter knows as "abst1." SM controls the rate of percolation, or depletion through a familiar exponential decay function,

$$SM2_i = SM1_{i-1}.exp^{(-\alpha_m.t_i)} \tag{2}$$

Where α_m is the monthly depletion factor for the m-th month which determines how fast the soil moisture is depleted. The water balance scheme further allows replenishment of SM by rainfall excess *F*:

$$SM1_i = SM2_i + F_i \tag{3}$$

Where F_i is defined as the difference between rainfall, initial abstraction I_{ai}, and runoff *QD*:

$$F_i = P_i - I_{ai} - QD_i \tag{4}$$

Note that Eqns. (2)-(4) essentially constitute a single-bucket soil moisture accounting scheme.

Though the manual by Matsumoto (1992) traces the origin of the soil moisture accounting scheme to Williams and LaSeur (1976), a more extensive literature review reveals that the scheme in fact closely resembles that proposed by Hawkins (1978), with the only distinction being that it considers percolation as the loss mechanism, rather than evapotranspiration (ET) as does the latter. Note that exponential decay seen in Eqn. (3) arises from the linear storage-loss relationships typically used in hydrologic literature (Kohler and Linsley, 1951; Budyko, 1961).

The baseflow under dry conditions is modeled using an exponential decay function.

$$QB2 = \mathrm{QB1.K}^{\mathrm{t_2 - t_1}} \tag{5}$$

where K is the recession constant. For daily simulation, TxRR adopts the SCS non-dimensional unit hydrograph (UH) to route the direct runoff QD. The UH parameter lag time T_l is estimated from the drainage area by an empirical relationship:

$$\mathrm{T_l = \beta A^{0.6}} \tag{6}$$

where the geomorphic coefficient β is treated as a calibration parameter. The time to peak T_p and base times T_b are computed separately by

$$T_p = 12 + T_l \tag{7}$$

$$T_b = 5.T_p \tag{8}$$

Calibration of TxRR typically involves the tuning of 15 parameters, namely maximum storage SMMAX, and 12 monthly soil moisture depletion constant $\alpha_{m.}$, baseflow recession constant K, and the coefficient β in Eqn (6), though parameters such as the abstraction to storage ratio abst1 can be adjusted when necessary.

TxRR offers a simple mechanism to transfer one of the parameters, SMMAX, between a pair of catchments. For a recipient catchment, the value of total soil water storage, SMMAX, is related to the calibrated value for a donor, gauged catchment via the following equation:

$$SMMAX_U = \left(\frac{CN_G}{CN_U}\right).SMMAX_G \tag{9}$$

where $SMMAX_U$ and $SMMAX_G$ are the maximum soil moisture storage for the ungauged (recipient) and gauged (donor) watersheds, respectively. And CN_U and CN_G are the corresponding curve number values. For the remaining three parameters, it is assumed that the calibrated values can be directly transcribed among watersheds without any modification. A caveat, however, is that the identification of donor-recipient catchment is not discussed in the manual and is left entirely to the user's discretion. For simplicity, in this study the donor catchment is chosen to be the one in the closest proximity to the recipient as measured by the distance between the watershed centroids.

The current operational version of TxRR was calibrated back in the late 1990s over gauged watersheds, but few details remain on the mechanism of calibration or the exact watersheds for which calibration was performed, and the calibrated parameter values were adapted to coastal, ungauged watersheds, but the original values for the gauged watersheds were not retained.

2.1.2. National Water Model (NWM)

The National Water Model (NWM) is an implementation of the WRF-Hydro system developed at National Centers for Atmospheric Research (NCAR; Gochis *et al.*, 2015, 2020). The model first came into real-time operation at NWS in 2015 (Cosgrove *et al.*, 2016); it has since undergone several upgrades and the current operational version is NWM 2.1. The NWM runs on a 1-km grid mesh to produce forecasts and analyses of streamflow, soil moisture, and snowpack for the entire Conterminous US (CONUS) and Alaska. The forecasts produced by NWM supplement the stage and discharge forecasts produced by NWS River Forecast Centers (RFCs) for designated forecast points in the US.

Fig. 2 illustrates the various model components and a schematic of the processes in producing forecast products. Unlike the TxRR and other lumped hydrologic models, the NWM is a gridded model whose major components include a gridded energy and water balance module, an embedded terrain routing module, and channel and reservoir routing modules. At present, the NWM uses the Noah-MP land surface model (Niu *et al.*, 2011; Yang *et al.*, 2011) for water balance calculations; a diffusive wave routing module for terrain routing; and the Muskingum-Cunge method for channel routing. The Noah-MP is an extension of the Noah land surface model with multiple suites of parameterization schemes (Chen and Dudhia, 2001; Ek *et al.*, 2003; Gan *et al.*, 2019). It employs a full-energy balance scheme that relies on meteorological observations, analysis, and surface conditions to calculate sensible and latent heat fluxes. Contrasting to some of the widely used land surface and hydrologic models, e.g., the one-bucket model by Manabe (1969), and the two-bucket Sacramento Soil Moisture Accounting (SAC-SMA; Burnash *et al.*, 1973), Noah-MP offers explicit representations of soil layers down to 2-m depth and flux exchange among the layers. In addition, it incorporates an unsaturated zone and allows it to interact with the groundwater aquifer. The model inherits some of the assumptions in the TOPMODEL (Beven and Kirkby, 1979), e.g., the exponential decrease in permeability with depth; and it represents infiltration excess runoff through a tunable parameter REFKDT proposed by Schaake *et al.* (1996).

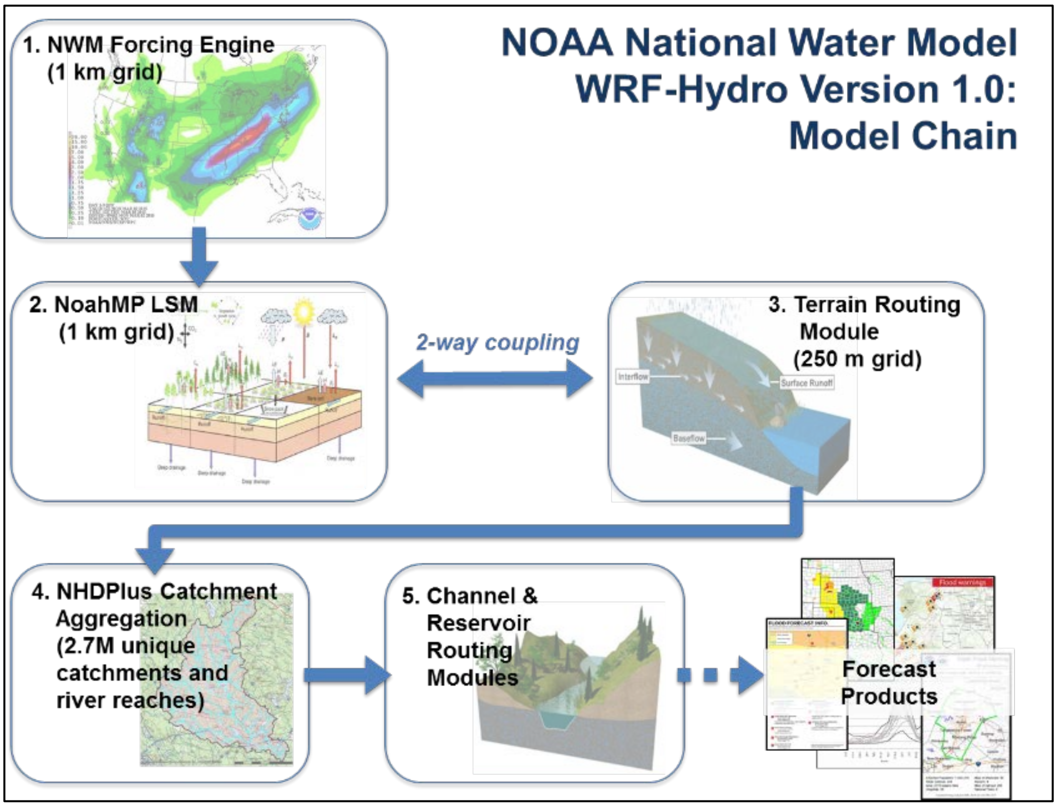

NOAA National Water Model
WRF-Hydro Version 1.0:
Model Chain
1. NWM Forcing Engine (1 km grid)
2. NoahMP LSM (1 km grid)
2-way coupling
3. Terrain Routing Module (250 m grid)
Interflow
Surface Runoff
Baseflow
4. NHDPlus Catchment Aggregation (2.7M unique catchments and river reaches)
5. Channel & Reservoir Routing Modules
Forecast Products
Flood warnings
FLOOD FORECAST INFO

Figure 2: Components of NWM and data flow diagram. Note the diagram applies to NWM 2.1 that is the current operational version (Source: NWS OWP).

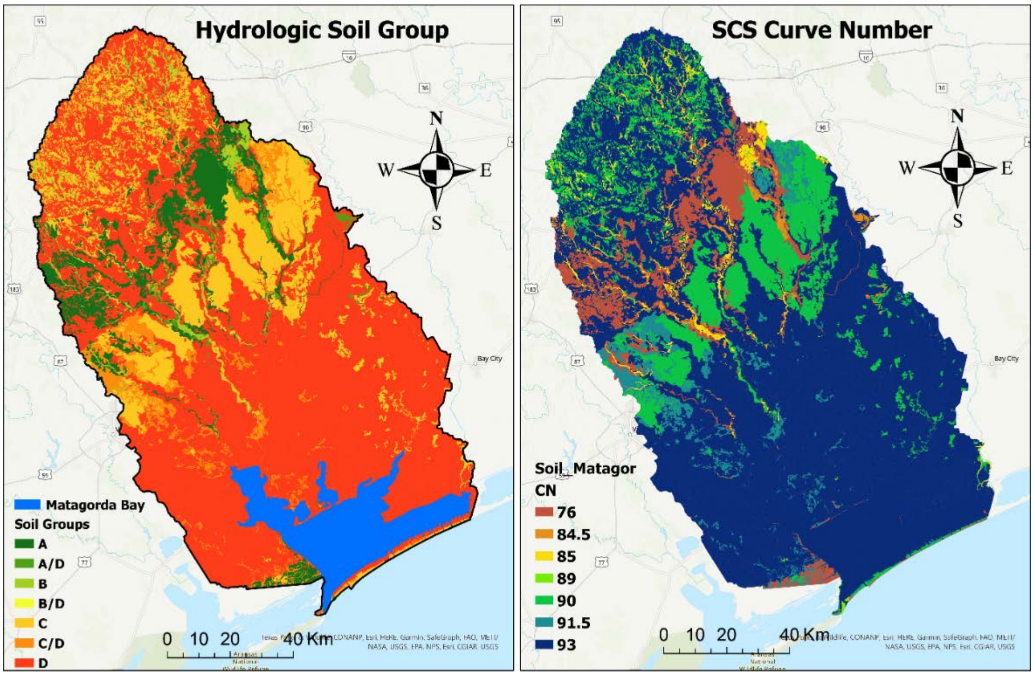
The terrain model of NWM simulates hillslope and groundwater flow processes on a 250-m mesh embedded in each 1-km cell: where the saturate subsurface flow is modeled using two-dimensional, quasi-steady state Boussinesq model and the hillslope routing is represented by either 1-D or 2-D diffusive wave routing (note: the 1-D routing scheme is the default for the operational version). NWM incorporates an exponential storage-discharge function to model baseflow that takes the form of groundwater release.

As an operational system, NWM is configured to run over different cycles, including analysis, short-range (0–15 hours), medium-range (0–10 days), and long-range (0–30 days) forecast. For real-time analysis and forecast, NWM performs a simple nudging operation to adjust streamflow prediction to match USGS streamflow observations. The model currently produces discharge analysis and forecasts for 2.7 million stream reaches included in the NHD-Plus database, including many coastal streams not being covered by NWS RFCs. This ability of NWM to produce streamflow analysis and prediction over ungauged reaches presents a key advantage over lumped models, and its coverage of coastal regions makes it a potentially useful source of information on freshwater inflows to the bays and estuaries. In addition, NWM's use of physics-based water and energy balance schemes may allow it to better represent aspects of runoff production that are not resolvable using TxRR.

In this study, we chose to use NWM 2.0 reanalysis, which is the latest reanalysis available. The reanalysis was created by running NWM version 2.0 over the 26-year window extending from 1993 to 2018 using forcing variables from the North America Land Information System – 2 (NLDAS-2; Xia *et al.*, 2012). These variables include precipitation, 2-m air temperature, radiation, wind speed, and relative humidity. The NWM 2.0 underwent limited calibration for 1451 selected gauging stations in the US (Nasab *et al.*, 2020). The calibration was performed for the period of 2008–2013, with 18 parameters adjusted to minimize a composite objective function that is based on Nash-Sutcliffe Efficiency (NSE; Nash and Sutcliffe, 1970), via the Dynamic Dimension Search algorithm documented in Tolson and Shoemaker (2007). Additional metrics, including the Multi-scale Objective Function (Kuzmin *et al.*, 2008), were monitored, but not directly used. The calibrated parameters are adapted to uncalibrated catchments that share similar hydroclimatic characteristics identified through cluster analysis (Nasab *et al.*, 2020).

2.2. Study Area and Methodology

The region of interest encompasses the upper, middle, and parts of the lower portions of Matagorda Basin along the central portion of the Texas coast. The Matagorda Basin is a collection of watersheds that drain to parts of the Lavaca-Colorado Estuary System, the second-largest estuarine system in Texas (Schoenbaechler *et al.*, 2011). The major streams in the study region include the Lavaca River and its tributaries, Navidad River, Sandy Creek, and Mustang Creek, with a combined drainage area of 11,825 km^2. Soil classification and land use information were acquired from the Multi-resolution Land Cover (MRLC; Wickham *et al.*, 2014), Texas Natural Resources Information System (TNRIS) and the NRCS SSURGO websites. Fig. 3 displays the soil classification-, curve number-, runoff-, and landuse-map over the region. It is evident that much of the lower portion of the basin is covered by agricultural lands, with predominately poorly drained, clayish soils. Along the middle portion of the basin lie patches of forests with well-drained and somewhat well-drained soils. Further upstream, forests are gradually replaced by agricultural lands and poorly drained soils resume domination. This gradation suggests that there is likely a rather wide variation in the runoff potential among these portions of the basin, with the lower portion being the most sensitive to rainfall input.



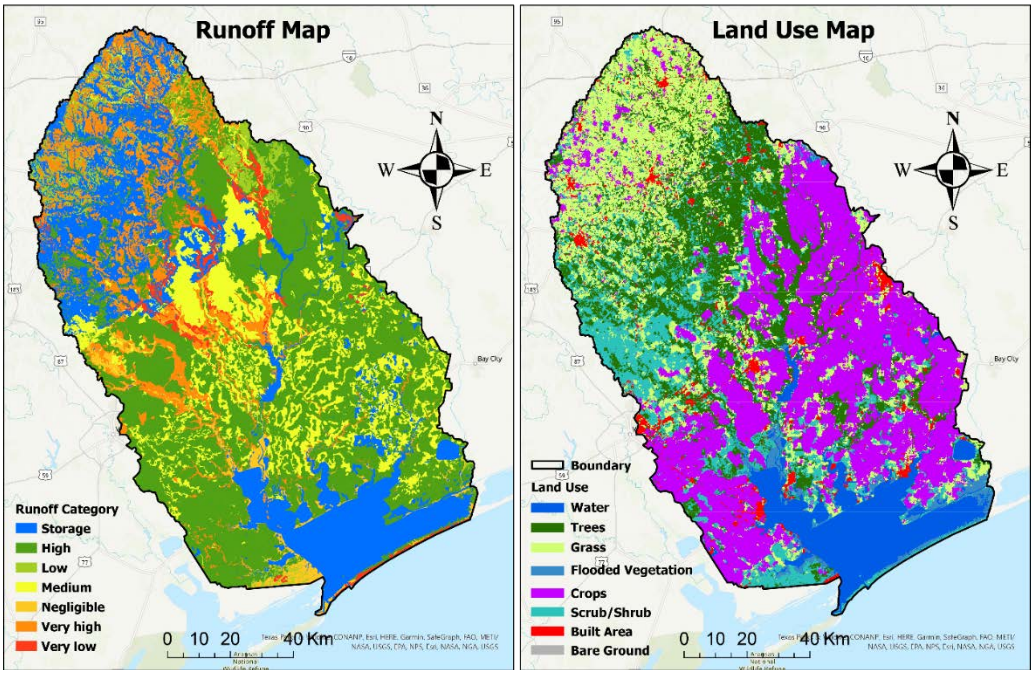


Figure 3: Soild classification map (a), Curve Number Map (b), Runoff Map (c), and land use map (d) of Matagorda Bay watersheds.

Nine USGS stations in the study region offer daily flow records as presented in Table 1. Six of these are situated along Lavaca River and its tributaries (Navidad River, Sandy Creek, and Mustang Creek), and three are on small coastal streams (Placedo Creek, Garcitas Creek, and Tres Palacios River). In order to differentiate the robustness of parameter transferability schemes for TxRR and NWM, watersheds draining to the nine stations are divided into two clusters (i.e., Group-1 and Group-2). The first group consists of five watersheds with TWDB identifiers, for which calibration was most likely done for TxRR (Table 1). Out of the five watersheds, NWM was calibrated for three, with in Standard Hydrologic Exchange Format (SHEF) codes of SBMT2, EDNT2 and NGCT2 (Nasab *et al.*, 2020). The remaining four watersheds are located along the east portion of the basin and are not included in the current TWDB's database of watersheds as shown in Fig. 4.

Table 1: Description of USGS stations having long-term streamflow observation located within the Lavaca River basin; the stations are divided into two groups; the light gray and light brown represent observations for Group 1 and Group 2 watersheds, respectively.

USGS ID	Station Name	TWDB ID	NWS SHEF ID*	Drainage Area [km^2]	Available Data Period
			Group 1		
08164300	Navidad River above Hallettsville	16007	SBMT2	860	From 1961-10-01 to date
08164000	Lavaca River near Edna	16005	EDNT2	2116	From 1938-08-13 to date
08162600	Tres Palacios River near Midfield	15020	MTPT2	404	From 1970-06-17 to date
08164600	Garcitas Creek near Inez	17020	NGCT2	277	From 1970-06-15 to date
08164800	Placedo Creek near Placedo	17040	PLPT2	184	From 1970-06-16 to date
			Group 2		
08164390	Navidad River at Strane Park near Edna	16014	LSNT2	1500	From 1996-10-01 to date

doi:10.20944/preprints202210.0415.v1

08164450	Sandy Creek near Ganado	16014	CODT2	749	From 1978-10-01 to date
08164503	West Mustang Creek near Ganado	16014	GNDT2	461	From 1977-10-01 to date
08164504	East Mustang Creek near Louise	16014	LMCT2	140	From 1996-10-01 to date

*SHEF stands for Standard Hydrologic Exchange Format used by National Weather Service.

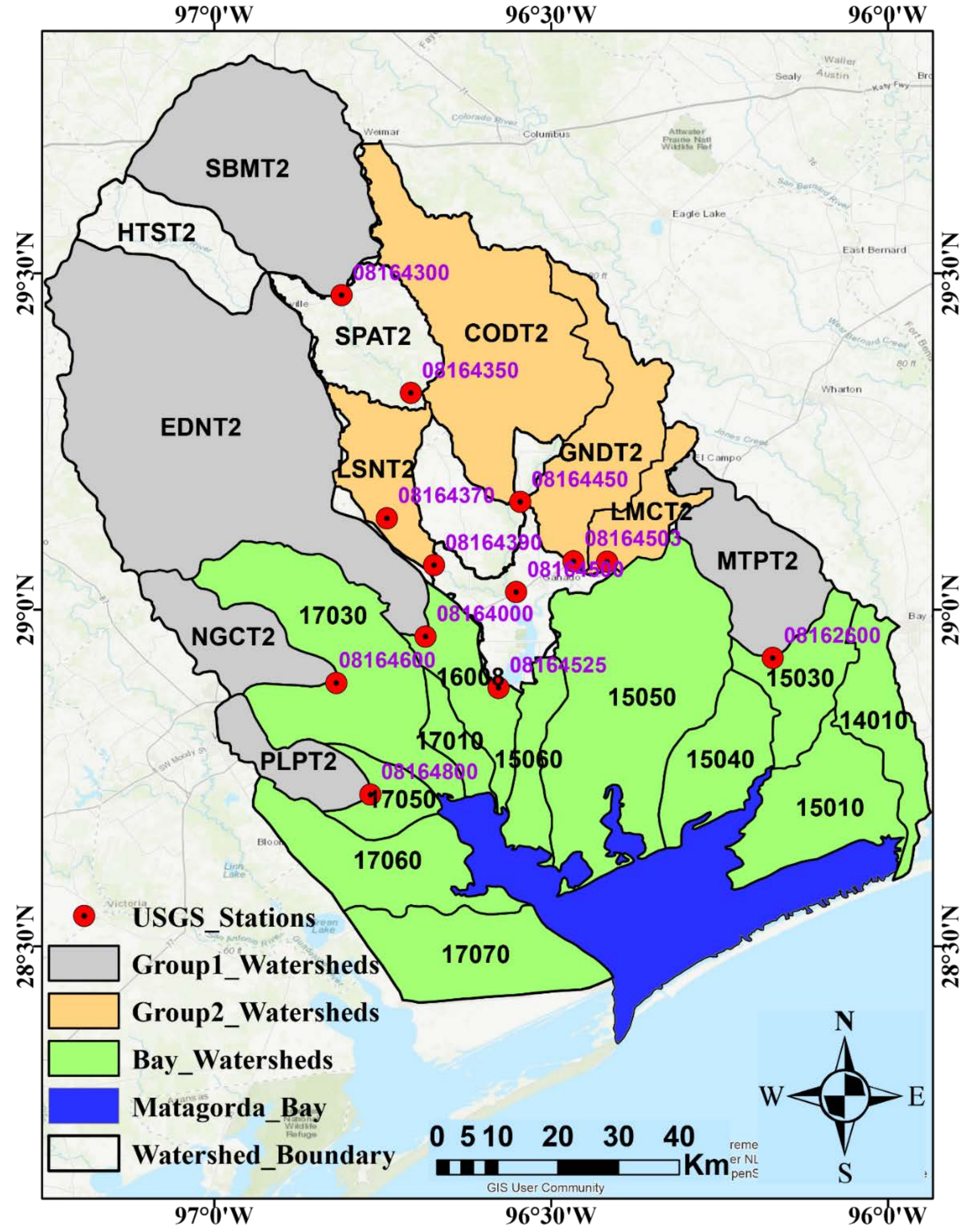


Figure 4: Cluster Map of Matagorda Bay watershed also known as Lavaca basin in Texas, USA. The light gray and light brown represent Group 1 and Group 2 watersheds, respectively. The red circles represent the USGS streamflow measuring stations, and the upstream catchments are named according to NWS SHEF ID; however, the inlet watersheds are numbered according to TWDB's watershed number.

Due to the aforementioned lack of record on the previous round of TxRR calibration, the TxRR parameter values maintained for coastal watersheds are reversely adapted to the five upstream watersheds in the first group. This adaptation entails establishing donor-recipient pairs using physical proximity between the watershed centroids. Among the four parameters, SMMAX was transferred from the recipient to donor using Eqn. (9), whereas the others are directly transcribed. The parameters over these five watersheds are then transferred to those in the group 2, again by establishing donor-recipient relationships via shortest distance between watershed centroids as presented in Table 2.

Table 2: Donor-Recipient basin pairs for bias correction and parameter transfer. The correction factor and model parameters are calculated for donor basins and transferred to recipient basin based on their physiological similarities.

Donor basins	Recipient basins
SBMT2	CODT2
EDNT2	LSNT2
MTPT2	LMCT2 and GNDT2

The TxRR operational daily runs are driven by a daily rainfall data set, which is created by interpolating daily Global Historical Climate Network (GHCN) gauging station reports to each watershed using the Thiessen Polygon method. NWM reanalysis, by contrast, was produced using the gridded NLDAS-2 forcing data set, whose precipitation grids are interpolated from reports by gauges in the NWS Cooperative Observer Program (COOP[1]), but using a different interpolation algorithm. In this algorithm, the gauges are divided among different topographic facets specified in the Parameter-elevation Regressions on Independent Slopes Model (PRISM; Daly *et al.*, 1994). Then the reports from each group of gauges are interpolated within the respective facet before being merged and mapped onto a 1/8-degree grid mesh, and the results are then disaggregated onto hourly increments using the hourly gauge data prior to 1995 or National Centers for Environmental Prediction (NCEP) Stage-II radar-based hourly precipitation analysis thereafter. Despite sharing a nearly identical gauge data set as the input, the TxRR rainfall series and the NLDAS-2 precipitation were found to differ substantially over the region – for some watersheds, the former is 10% greater than the latter on a cumulative basis. In order to rule out differences stemming from those in forcing inputs, in this study the NLDAS-2 precipitation data set is used to create mean areal precipitation data set to serve as the input to TxRR. TxRR is run for each of the nine watersheds using this forcing data set as input for 1993-2017. The performance differentials between the two models are examined from four perspectives: with the respective foci on 1) relative accuracy of model simulations and dependence on temporal scales; 2) potential impacts of differences in calibration time windows; 3) ability of models to reproduce distributions of flow regimes; and 4) ability of models to capture flow series during selected flooding and low flow events. In addition to these, evaluation of model performance for Group 2 watersheds helps illuminate the efficacy of the parameter transfer scheme of each model.

As the study region is situated in a relatively humid region (with aridity index>0.65), the role of net incoming solar radiation is expected to be dominant in regulating ET per Budyko's hypothesis (Budyko, 1974). While TxRR accounts to an extent seasonal variation of ET through the use of monthly losses, the invariance of losses across years ignores interannual variations in the meteorological conditions that could impact evaporation at soil surface and transpiration at the rootzone, which may in turn alter the dynamics of soilwater storage. We hypothesize that the NWM, through its use of explicit energy and water balance schemes, offers enhanced realisism in capturing this interannual variations in soil moisture dynamics. To test this hypothesis, model performance comparisons are performed over daily, monthly, and annual time windows. Conventional metrics employed

[1] GHCN is a subset of COOP network. Most COOP stations in the region are in the GHCN network.

doi:10.20944/preprints202210.0415.v1

for judging model performance include percentage bias (PB), Pearson's correlation (R), and root mean squared error (RMSE). These metrics are defined as follows:

$$PB = \frac{\sum_{i=1}^{n}(Q_{sim,\mathrm{i}} - Q_{obs,i})}{\sum_{i=1}^{n} Q_{obs,i}} \times 100 \tag{10}$$

$$\mathrm{R} = \frac{\sum_{i=1}^{n}(Q_{obs,i} - \bar{Q}_{obs,i})(Q_{sim,i} - \bar{Q}_{sim,i})}{\sqrt{\sum_{i=1}^{n}(Q_{obs,i} - \bar{Q}_{obs,i})^2}\ \sqrt{\sum_{i=1}^{n}(Q_{sim,i} - \bar{Q}_{sim,i})^2}} \tag{11}$$

$$\mathrm{RMSE} = \sqrt{\frac{\sum_{i=1}^{n}\left(Q_{sim,i} - Q_{obs,i}\right)^2}{n}} \tag{12}$$

where $Q_{obs,i}$ and $Q_{sim,i}$ are observed and simulated streamflow, respectively; and n is the number of records in the time series.

TxRR was calibrated more than 20 years ago, and therefore there is a distinct possibility that the model performance has degraded over time. NWM was calibrated for the period of 2008–2013, in which dry conditions prevailed across the study region. The impacts of differential calibration time windows require special attention. To this end, this study employs a sliding window-based comparison wherein the accuracy of simulated flows from each model is calculated and compared separately for each of four overlapping 10-year windows, i.e., 1993-2002, 1998-2007, 2003-2012, and 2008-2017.

The performance of the two models is closely scrutinized for two major flooding events and over a rainfall episode that took place in a severe drought, with the aim of discerning their differential abilities to capture rainfall-runoff processes under abnormal conditions. The two flooding events occurred during from 15–23 October 1994 and during Hurricane Harvey from August 25–September 5 of 2017. The third episode spanned September and October of 2011 and was embedded in the severe drought of 2011–2014.

Finally, the differential efficacy of parameter transfer schemes of TxRR and NWM is examined. For this purpose, TxRR and NWM simulations are compared for four additional gauged watersheds in Matagorda Basin that are not included in the database of TWDB (Group 2 watersheds in Table 1). All these watersheds drain to the Lavaca River; the drainage areas of the watersheds range from 140 km^2 to 1500 km^2 and with about 28 years of flow records. To establish TxRR parameters for each watershed, a donor watershed from Group 1 is determined using the shortest distance between watershed centroids. Then, the SMMAX value is transferred from the donor to the recipient via Eqn (9). Note that NWM has not undergone calibration for any of these Group 2 watersheds; though, as indicated earlier, the calibrated NWM parameter values have been adapted to these watersheds most likely from the three calibrated watersheds in Group 1.

3. RESULTS

3.1. Assessment of Overall Simulation Accuracy

Table 3 summarizes the relative performance of the two models for the first group of watersheds over the entire period (1993–2017). In terms of percentage bias, NWM and TxRR perform comparably, with the former producing lower biases for a slight majority of the basins (three out of five). It is interesting to note that the direction of bias is largely consistent between the two models. For SBMT2, the northmost watershed, simulations from both models are positively biased. For the remaining basins, the biases for both set of simulations tend to be negative. The NWM simulation features near neutral bias for the largest watershed EDNT2 (Lavaca River near Edna), but it exhibits glaringly negative bias near the bay at MTPT2 (Tres Palacios River near Midfield).

doi:10.20944/preprints202210.0415.v1

Table 3: Validation statistics of daily, monthly, and yearly flow for Group-1 watersheds produced by TxRR and NWM for the study period (1993-2017). Dark green, light green, and gold color represent outperformance of the respective model based on Pearson's correlations, RMSE, and percent biases, respectively.

Daily						
SHEF ID	R		RMSE [cms]		%BIAS	
	NWM	TxRR	NWM	TxRR	NWM	TxRR
EDNT2	0.67	0.68	54	55	-3	-27
SBMT2	0.64	0.67	22	21	24	30
NGCT2	0.66	0.74	9	8	-31	-16
MTPT2	0.71	0.82	13	10	-53	4
PLPT2	0.65	0.65	6	6	-5	-18
Monthly						
SHEF ID	R		RMSE [cms]		%BIAS	
	NWM	TxRR	NWM	TxRR	NWM	TxRR
EDNT2	0.81	0.83	16	16	-3	-27
SBMT2	0.74	0.79	6	6	25	30
NGCT2	0.86	0.86	3	2	-31	-16
MTPT2	0.76	0.88	5	3	-53	4
PLPT2	0.78	0.75	2	2	-5	-18
Yearly						
SHEF ID	R		RMSE [cms]		%BIAS	
	NWM	TxRR	NWM	TxRR	NWM	TxRR
EDNT2	0.93	0.84	4	7	-3	-27
SBMT2	0.89	0.77	2	3	25	30
NGCT2	0.93	0.91	1	1	-31	-16
MTPT2	0.88	0.91	3	1	-53	4
PLPT2	0.85	0.83	1	1	-5	-18

A notable observation is that the relative performance of the models, as judged by correlation and RMSE, displays a striking dependence on temporal scale. Both at daily and monthly scale, TxRR outperforms NWM in terms of correlation for all five watersheds, but at yearly scale the results become mixed with TxRR underperforming NWM for all watersheds except MTPT2. In terms of RMSE, NWM outperforms for only one out of five watersheds at the daily scale and demonstrates mixed results at monthly scale, possibly a consequence of narrower biases; however, at the yearly scale NWM outperforms for all watersheds except MTPT2. While alternative explanations cannot be ruled out at this point, the most likely determinant of this scale-dependent performance of NWM is its ability to account for temporal variations in meteorological and land surface conditions in calculating evaporative fluxes. As mentioned earlier, Budyko's insight is that in humid regions, it is the net energy flux that determines ET and in turn modulates soil moisture dynamics. Relative to rainfall, ET exerts influence on runoff dynamics over longer time scales. While calibration of TxRR seasonal loss rates helps represent the seasonal trends of ET to an extent, it appears this alone is incapable of capturing the slower modes in soil dynamics driven by interannual variations in net radiative fluxes.

Fig. 5 shows the time series of annual observed versus simulated flows by TxRR and NWM for each of the Group 1 watershed. The performance differential exhibits wide variations among the watersheds. Notable observations include the follows. First, for three watersheds, namely EDNT2, NGCT2 and PLPT2, TxRR is unable to capture the annual flows during some of wetter years. For EDNT2, the negative bias in TxRR is evident in the earlier years (up to 2005), whereas NWM performs consistently better. For NGCT2, the flows for the wetter years are consistently underestimated by both models, though the bias is more severe for NWM. For MTPT2, NWM simulation exhibits persistent biases throughout the entire period.

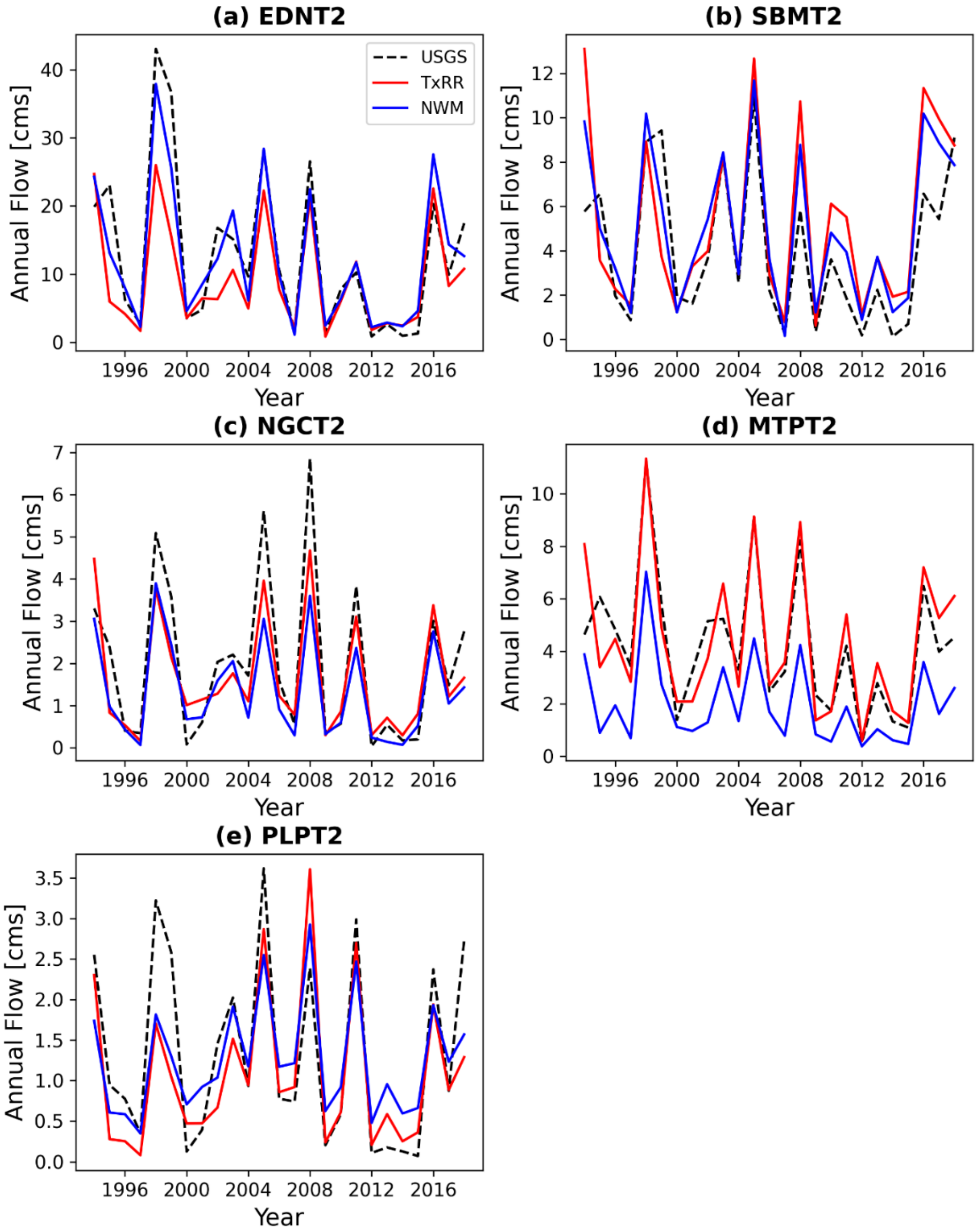
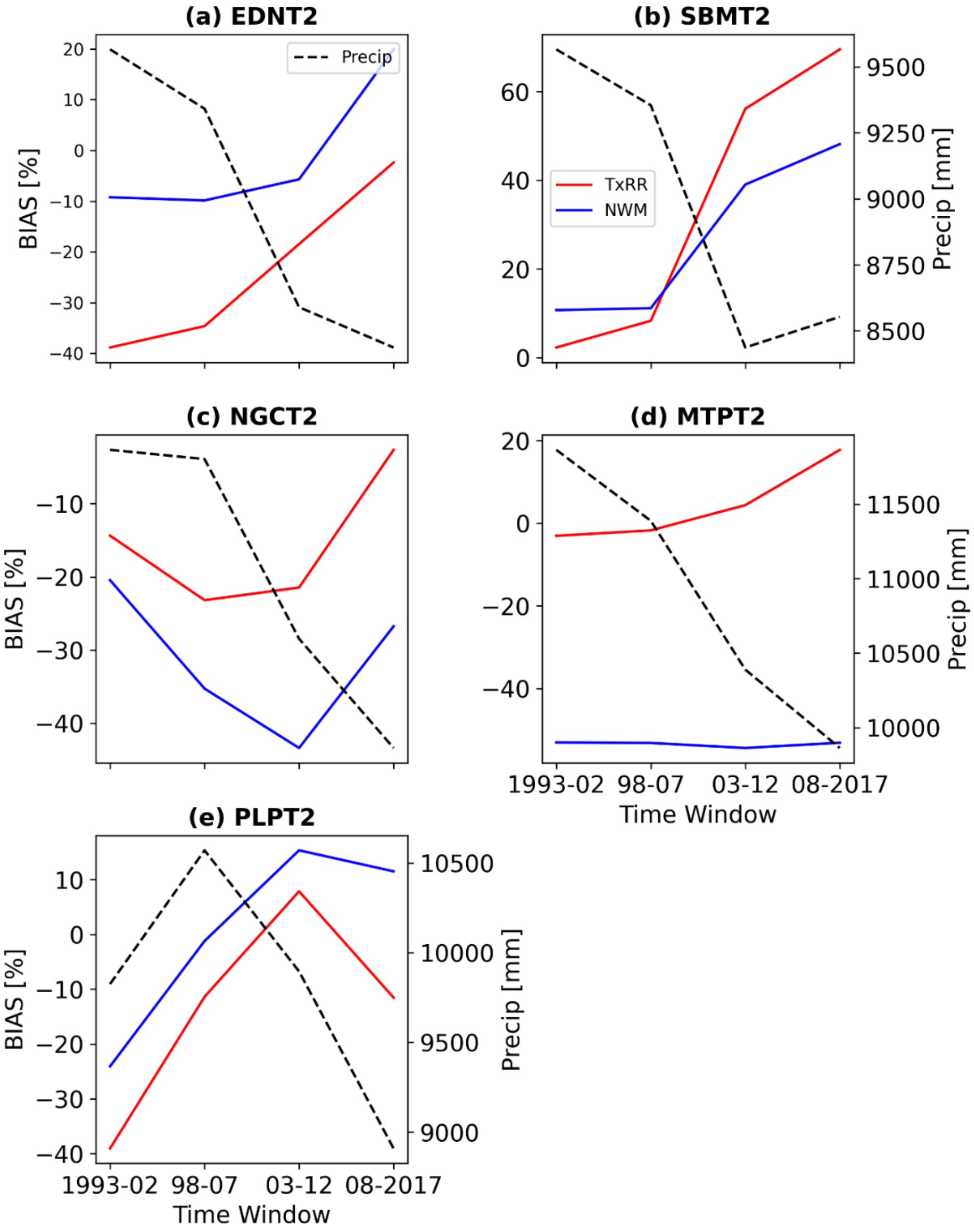


Figure 5: Comparison of annual mean flow produced by the TxRR and NWM models for five USGS stations located at the outlets of Group-1 watersheds. .

The performance of the two models at monthly scale over the sliding windows are shown in Figs. 6 and 7 in terms of bias and correlation, respectively. For all watersheds, there is a discernible downward trend in precipitation from the 1990s to the recent decade, which reflects the presence of the drought of 2011–2014. For two watersheds in the north, namely SBMT2 and EDNT2, biases in both TxRR and NWM exhibit clear, upward trends. Trends in biases are less clear for NGCT2 and PLPT2; for MTPT2, while TxRR simulation exhibits a monotonic upward trend, NWM simulation remains severely biased throughout the entire period. The relative skills of the two models largely mirror those computed on an aggregate basis, yet some features are worth noting. For example, for SBMT2, TxRR simulation is nearly bias neutral for the beginning of the period, but this bias grows increasingly severe in time, while the bias in NWM simulation also magnifies with time, the rate of increase is lower and this results in the better overall bias of NWM. Similar degradation in TxRR bias is also observed in MTPT2. For EDNT2 and PLPT2, the TxRR

doi:10.20944/preprints202210.0415.v1

simulation exhibits severe negative bias for the earliest window, which may reflect discrepancies in runoff dynamics during the window of calibration versus the window of validation. The USGS station Lavaca River at Edna maintains daily flow records that go back to 1938, and it is likely that the calibration of TxRR was done using the entirety of the record and reflects the runoff over "normal" conditions. The wet condition in the 1990s was possibly an aberration from the historical norm and this departure gives to the severe negative bias. This, however, does not explain the wide variations in the biases for the early time window among other watersheds for which daily records started from 1970. Another plausible explanation is related to the inconsistent biases in the precipitation input – precipitation amounts from the NLDAS-2 data set, as indicated earlier, are broadly lower (by as much as 10%) than those from the TWDB interpolated product used in calibrating the TxRR. Therefore, replacing this data set with NLDAS-2 which exhibits lower precipitation amounts almost certainly would reduce simulated runoff volumes, even if the model had been perfectly calibrated.



doi:10.20944/preprints202210.0415.v1

Figure 6: Comparisons of percent bias (PB) of monthly flow using 10-year sliding window between simulated flow produced by the TxRR and NWM for five USGS stations located at the outlets of Group-1 watersheds. .

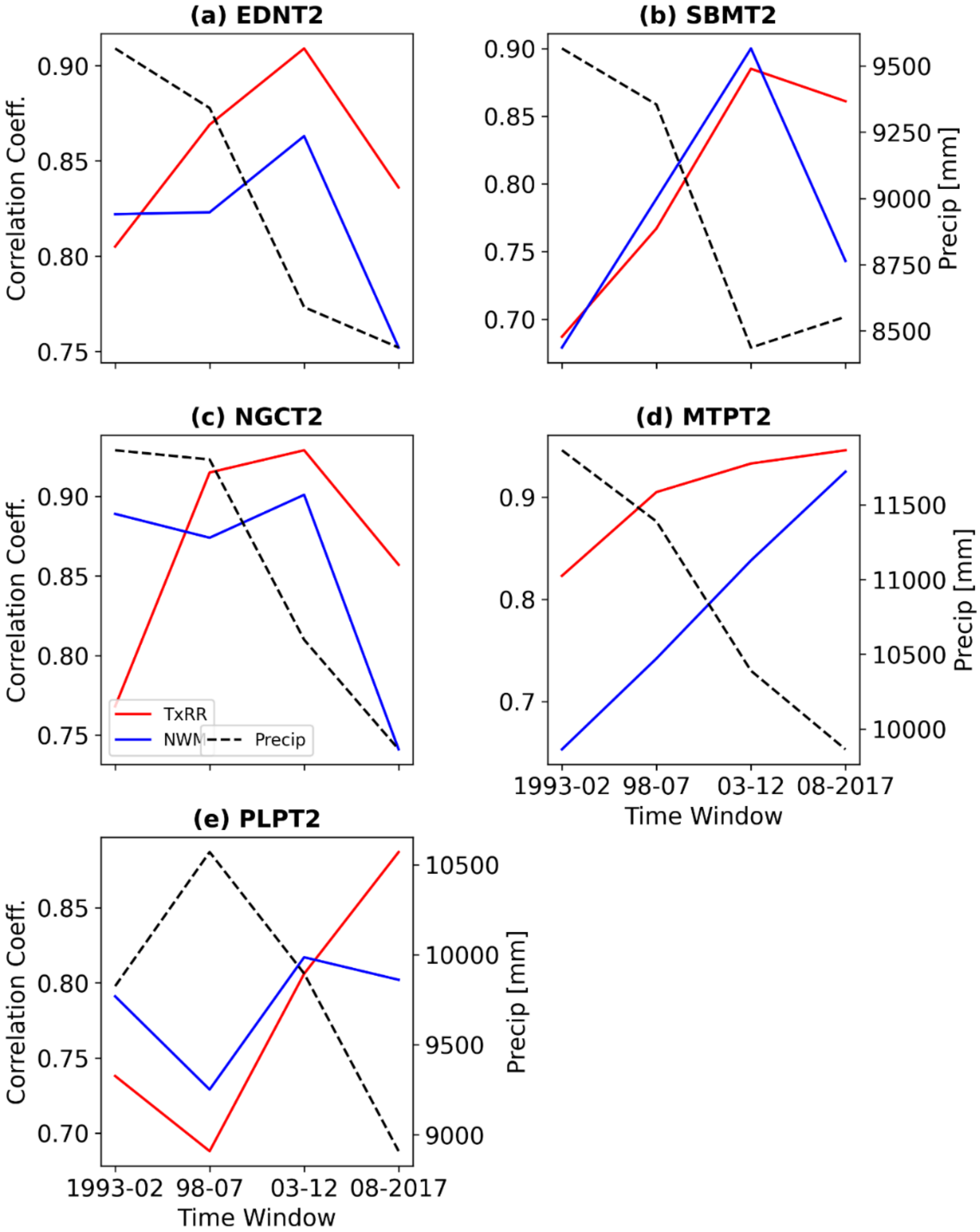


Figure 7: Comparisons of correlation coefficient of monthly flow using 10-year sliding window produced by the TxRR and NWM for five USGS stations located at the outlets of Group-1 watersheds. .

The correlation of monthly flow for the sliding windows is shown in Fig. 7. Again, the results are mixed across watersheds. For three watersheds, EDNT2, NGCT2, and MTPT2, TxRR simulation features higher correlation nearly across all the time windows. For MTPT2, correlation for both TxRR and NWM exhibits a monotonic rising trend, whereas for other watersheds trends are less obvious. As indicated earlier, NWM was calibrated for three of the five watersheds (SBMT2, EDNT2, and NGCT2) over the window of 2008–2013. Curiously, correlation for NWM simulation remains lower for two of the watersheds over the later sliding windows that overlap with the period used in calibration. The calibration of NWM, apparently, was not able to alter the skills of its simulations relative to those by TxRR either on a time aggregate basis or over the window where NWM was calibrated.

3.2. Case Studies

3.2.1. Flood of October 1994

The first flooding event occurred during 15–23 October 1994 as a result of a series of strong thunderstorms across the south-southeast Texas coast. The genesis of these storms can be traced to a curious encounter between the remnant of Hurricane Rosa (a hurricane that made landfall over the Pacific coast of Mexico) that was advected eastward by the westerlies, and a southward propagating low that originated from the Rockies. The merger of the two systems ignited powerful disturbances that were fueled by an ample supply of low-level moisture from the Gulf and unleashed heavy rainfall and widespread flooding. During this event, a rain gauge near the Lavaca River recorded 457 mm of rain; the peak water level in Lavaca River at the USGS station near Edna (08164000) exceeded the record flood stage set in 1936; and the peak discharge nearly doubled that for the latter event.

The spatial distribution of storm total rainfall is shown in Fig. 8, where it appears that much of the rain fell over the central and northern portion of the watershed and the accumulation diminished towards the bay. The hyetographs and hydrographs for this event are shown in Fig. 9. It is evident that both models grossly underpredict the flow volumes, and this underprediction is uniform across the five watersheds. For EDNT2, simulated daily flows from both models peak below 566 m^3s^{-1}, which is less than 20% of the observed daily peak (~3400 m^3s^{-1}). Similar, severe underprediction is observed for SBMT2 and PLPT2. Between the two models, NWM fares somewhat better for EDNT2 and NGCT2, but underperforms for the rest. The comparison of the mean areal precipitation product from TWDB versus that based on NLDAS reveals that the accumulation based on the former is more than twice that from the latter (390 vs. 170 mm). Such a wide difference points to a possible error in either NLDAS-2 data feed or processing algorithm. Another intriguing feature is that, for a majority of the watersheds, hydrographs based on TxRR simulations exhibit an earlier rise. This is possibly a reflection of the model's overrepresentation of soil moisture condition and runoff potential at the onset of the storm.

doi:10.20944/preprints202210.0415.v1

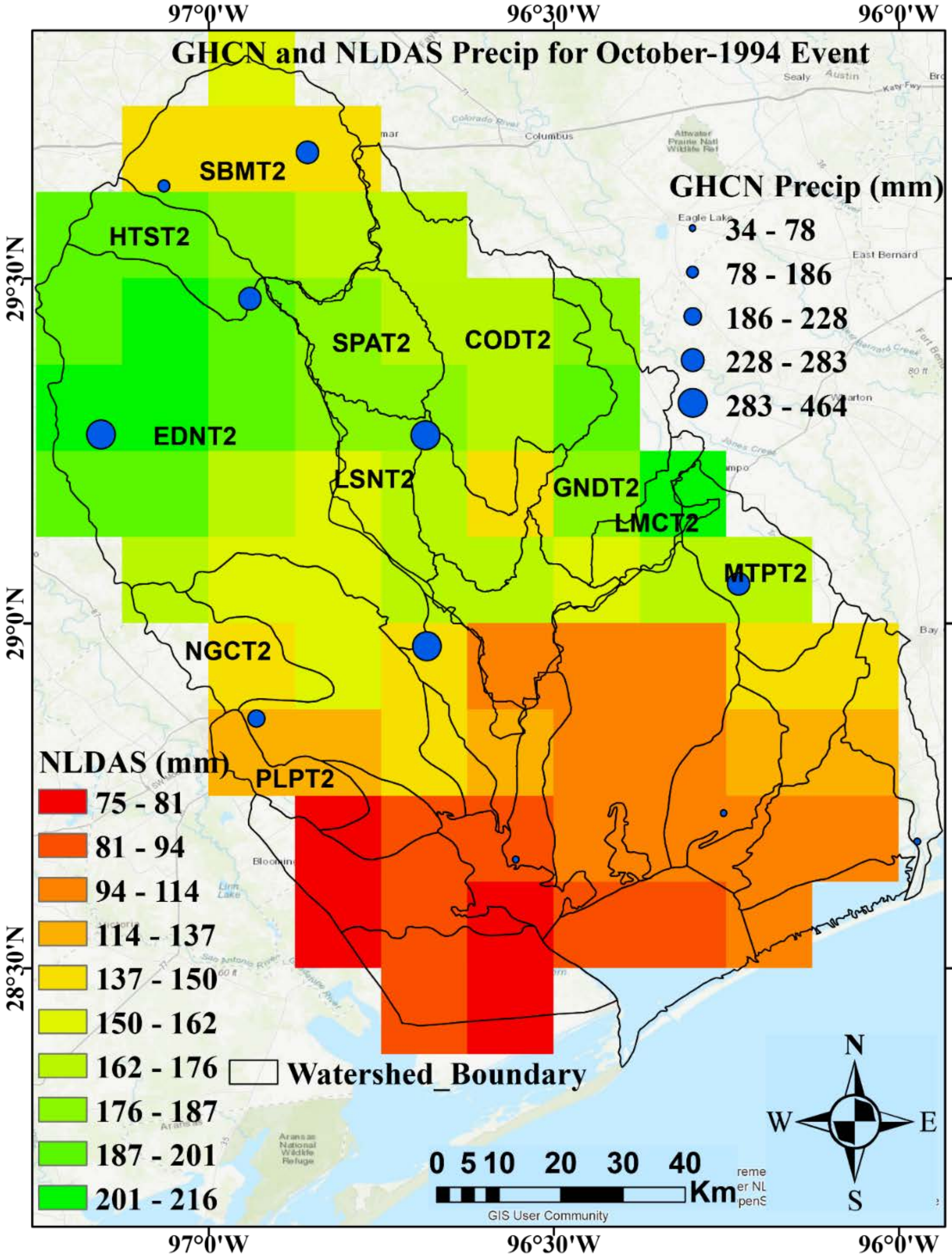


Figure 8: Rainfall accumulations for the October 1994 flood event based on the NLDAS-2 product. Superimposed are GHCN stations with the size scaled according to the recorded rainfall amount. .

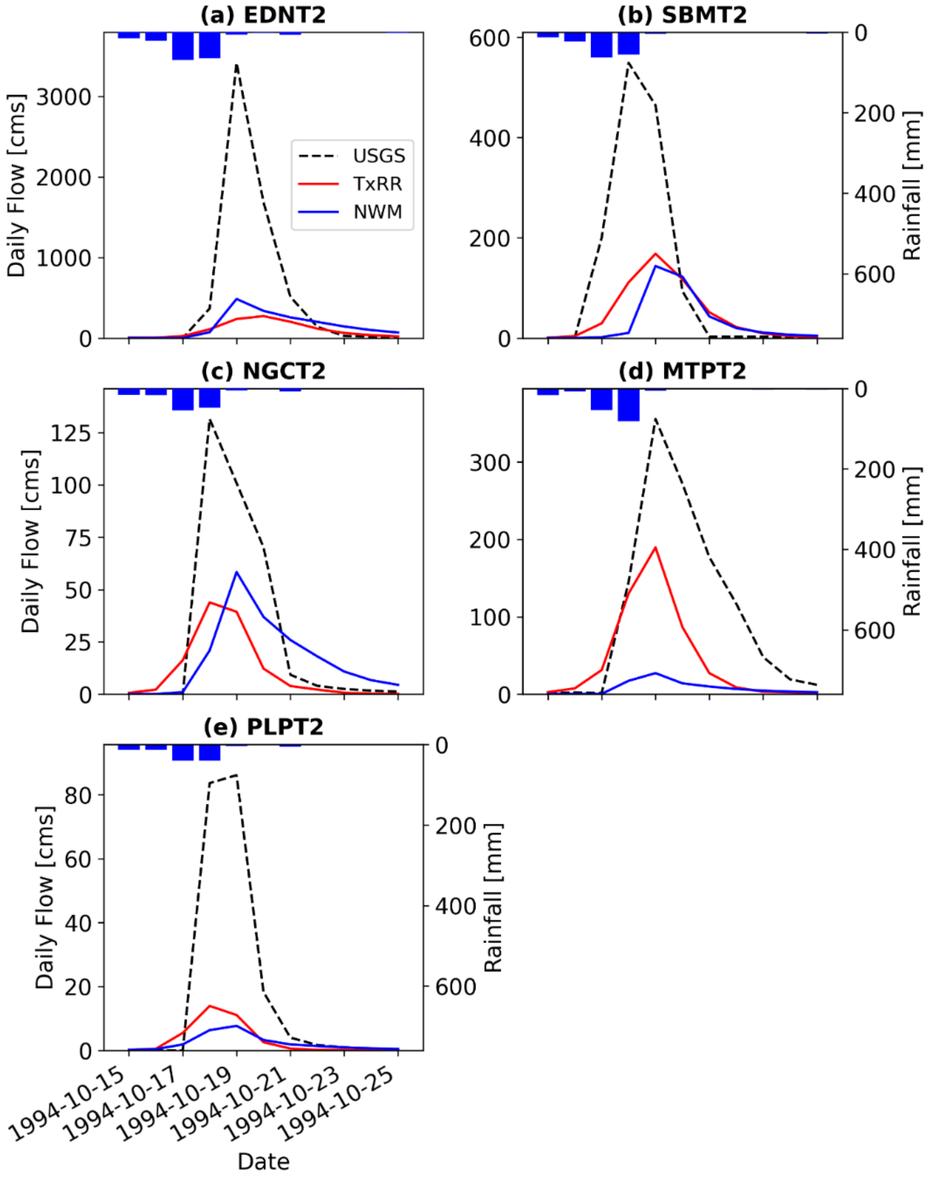


Figure 9: Comparisons of simulated flow produced by TxRR and NWM models and USGS observations for an extremely high flow event occurred in 1994, also known as October event.

3.2.2. Hurricane Harvey

The second flood event occurred as a result of Hurricane Harvey that took place in August 25–31, 2017. Harvey made landfall on August 26 near San Jose Island, and produced a sizable storm surge in Matagorda Bay. Its subsequent stalling brought torrential rain along the Southeast Texas coast for three days. The Matagorda Basin is further away from the rainfall center, and much of the region was under the "rain shadow": only about 170mm of rainfall fell over the central-lower portion of the Lavaca River Basin, whereas much higher accumulation was seen over the northern headwaters (~380mm; Fig. 10). Despite the relatively unimpressive rainfall input, the USGS station on Lavaca River near Edna reached the third highest stage in history, and at least one station near the bay (PLPT2, e.g., Placedo Creek at Placedo, USGS ID 08164800) also were near record stage due to a combination of rainfall and upstream propagating surge.

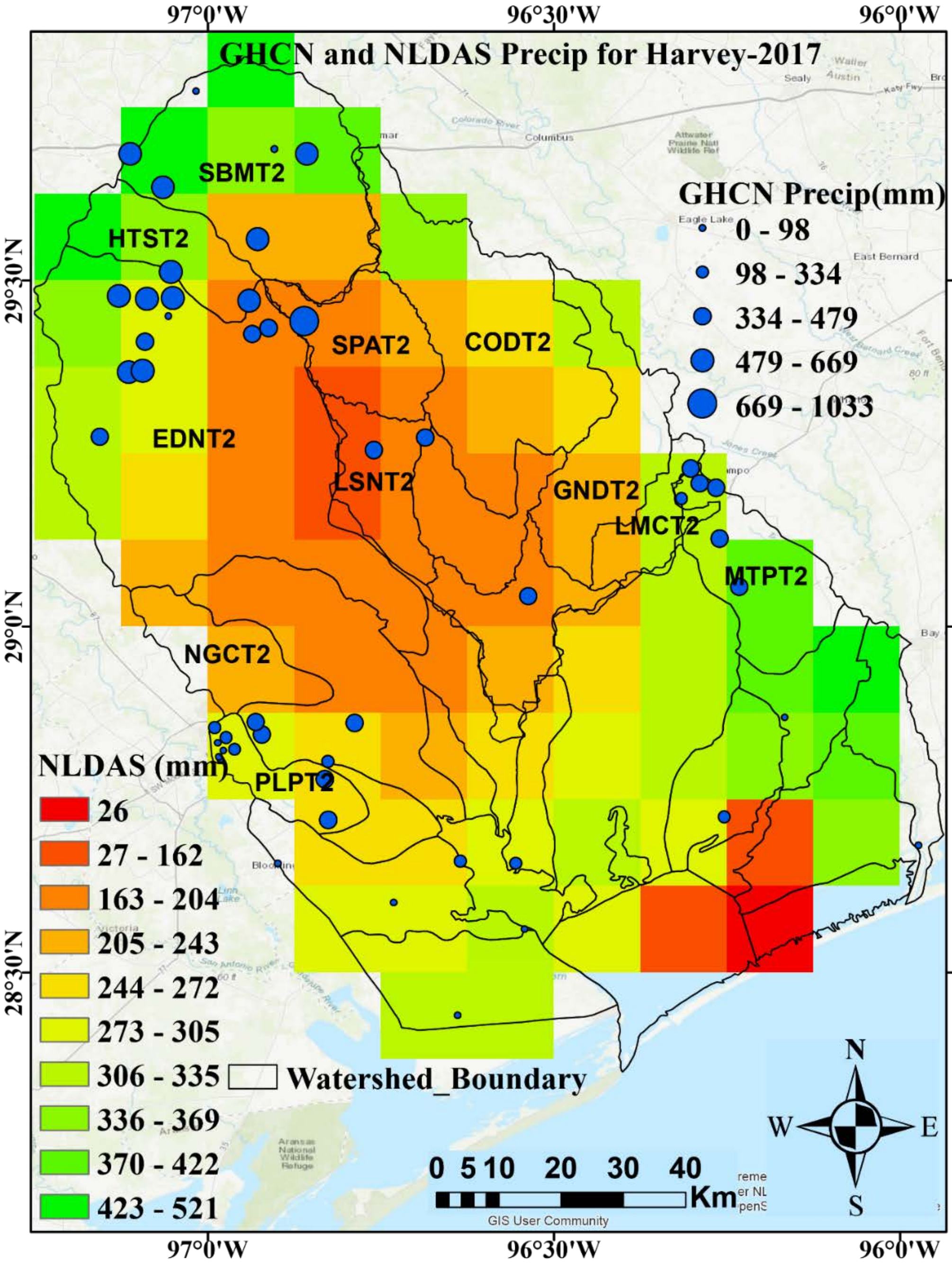


Figure 10: Rainfall accumulations for Hurricane Harvey episode based on the NLDAS-2 product. Superimposed are GHCN stations with the size scaled according to the recorded rainfall amount. .

Fig. 11 displays the daily hyetographs and hydrographs for the five Group 1 watersheds. Out of the five watersheds, four feature severe, negative biases in the simulated flows by both models. The severity of the negative bias varies among watersheds and between models. For EDNT2, the observed peak daily flow exceeds 1700 m^3s^{-1}, whereas simulated peak from either model is under 566 m^3s^{-1}. There appears to be a slight, but visible tendency for the bias to improve near the coast. For PLPT2 and MTPT2, NWM accurately reproduces the peaks; TxRR consistently underpredicts flow for PLPT2, but overpredicts the daily peak for MTPT2. The underprediction is at least in part attributed to issues in rainfall inputs. A comparison of NLDAS-2 and TWDB interpolated gauge products again points to a possible negative bias in the former product for this event. For

EDNT2, rainfall accumulation based on the latter product is about 50% higher than that from NLDAS-2.

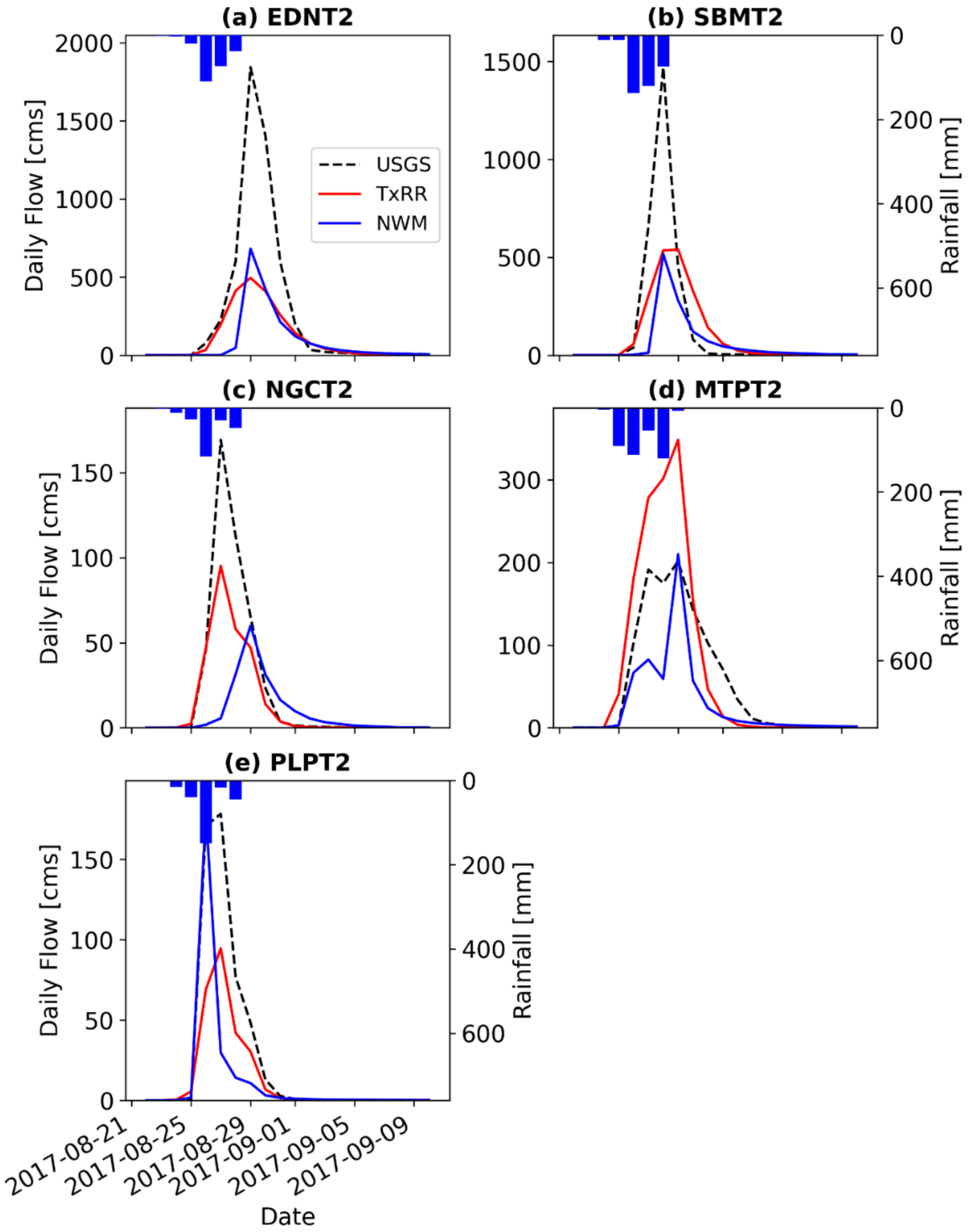


Figure 11: Comparisons of simulated flow produced by TxRR and NWM models and USGS observations for an extremely high flow event occurred in August 2017, also known as the Harvey event.

A salient feature in Fig. 11 is that NWM has difficulties capturing the early rise of the hydrographs for three watersheds in the upper and middle portion of the basin (EDNT2, SBMT2, and NGCT2), whereas TxRR does not. The contrast is particularly sharp for NGCT2. This difference in hydrograph timing is reminiscent of that for the October 1994 event, where the NWM produces hydrographs that consistently lag behind those based on TxRR, though for the present event it is the TxRR hydrographs that better resolve the timing of the observations. The mechanisms that underpin the timing difference are yet to be determined, but contrasts in the representation of soil moisture states at the onset of the event most likely play a critical role.

There is a remotely plausible thesis that the earlier portion of the observed discharge hydrograph for some of the stations was inflated by storm surge and associated backwater effects which preceded the arrival of the flood crest. This thesis, however, is challenged by the fact that NGCT2 (Garcitas Creek near Inez), the station for which NWM-observation discrepancy is the most pronounced, features steep bed elevation gradient (~10m above MSL at zero height versus 3m above MSL at EDNT2), where the impact of storm surge was likely subdued. Further research is required to discern the impacts of the surge on the daily flows during the event.

3.2.3. Drought Episode of 2011

The drought episode chosen for this study spanned September and October of 2011. Much of the early 2011 was abnormally dry with Palmer Drought Index approaching historical record. During spring and early fall of 2011, flows along nearly all streams in the Matagorda Basin largely vanished. A storm event occurred around October 10 that produced substantial rainfall accumulations across the region (50-100 mm), and much of the rainfall fell along the coast (Fig. 12). This storm, however, failed to elicit meaningful runoff response in most of the streams. The streamflow in Lavaca River, for example, was less than 0.2 m^3s^{-1} after the rainfall. Even for streams near the coast where higher accumulations were seen, e.g. PLPT2 (Placedo Creek), barely any runoff was produced during the storm.

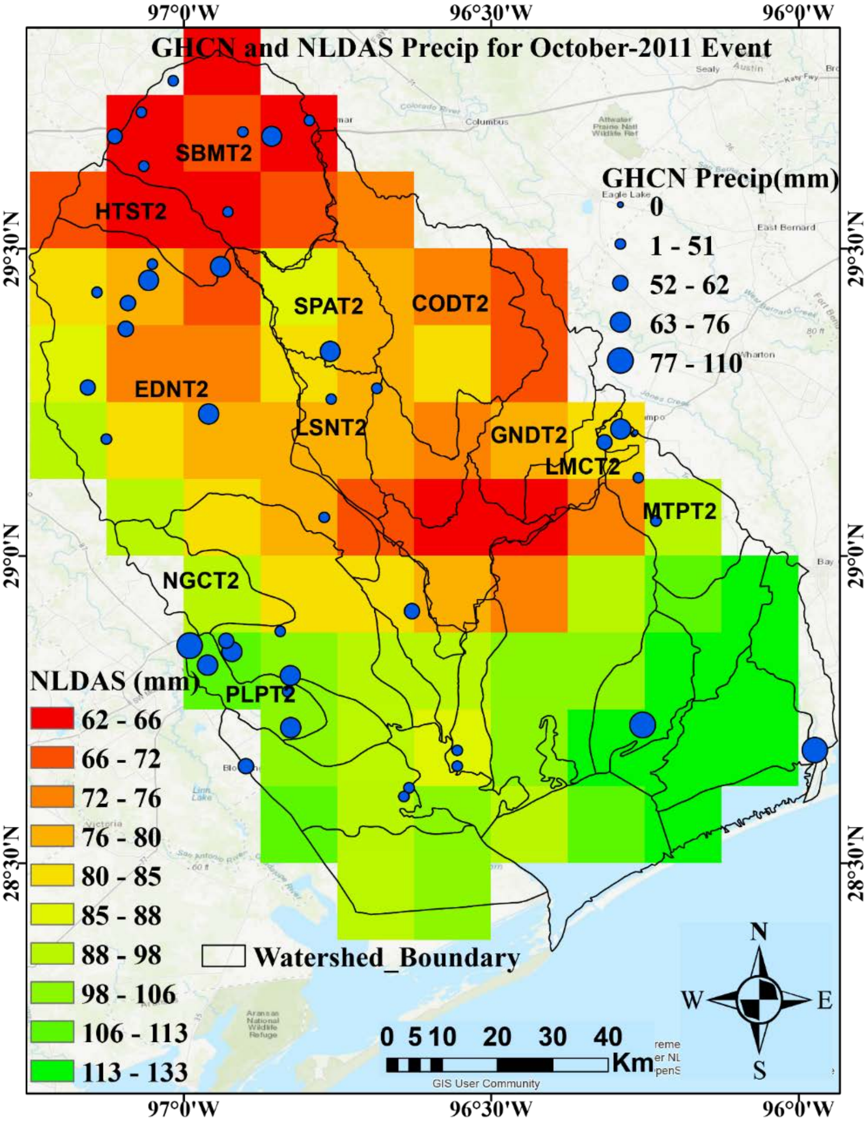


Figure 12: Rainfall accumulations for the October 2011 storm event based on the NLDAS-2 product. Superimposed are GHCN stations with the size scaled according to the recorded rainfall amount. .

The rainfall and streamflow series for this episode are shown in Fig. 13. The performance contrast between two models is stark. TxRR produces significant flows in response to the rain for all watersheds, whereas NWM fares better by producing little to no flow for three of the watersheds. For two coastal watersheds, namely PLPT2 and MTPT2, both models overpredict – for MTPT2 the TxRR simulation is more severely biased, whereas the opposite is true for PLPT2. This model performance differential can be possibly explained by a combination of differences in mechanistic representations of runoff generation, differing time windows employed in model calibration, and potentially contrasts in calibration strategies. To elaborate, TxRR uses time-invariant exponents derived from model calibration, and the magnitude of loss calculated using such exponents may be too modest to attain the abnormal dry condition and large infiltration losses during the 2011. By contrast, NWM incorporates a physically realistic representation of the ET process that would allow for more severe depletion of soil moisture contents during a drought, and this in turn allows for higher infiltration losses. In addition, the drought episode fell within the window during which NWM was calibrated (2008–2013), a window that has a significant overlap with the drought of 2010–2014. It is also possible that the newer round of calibration performed for the NWM 2.0 has a primary effect of reducing false streamflow. These, collectively, result in the superior performance of NWM 2.0 reanalysis as observed herein.

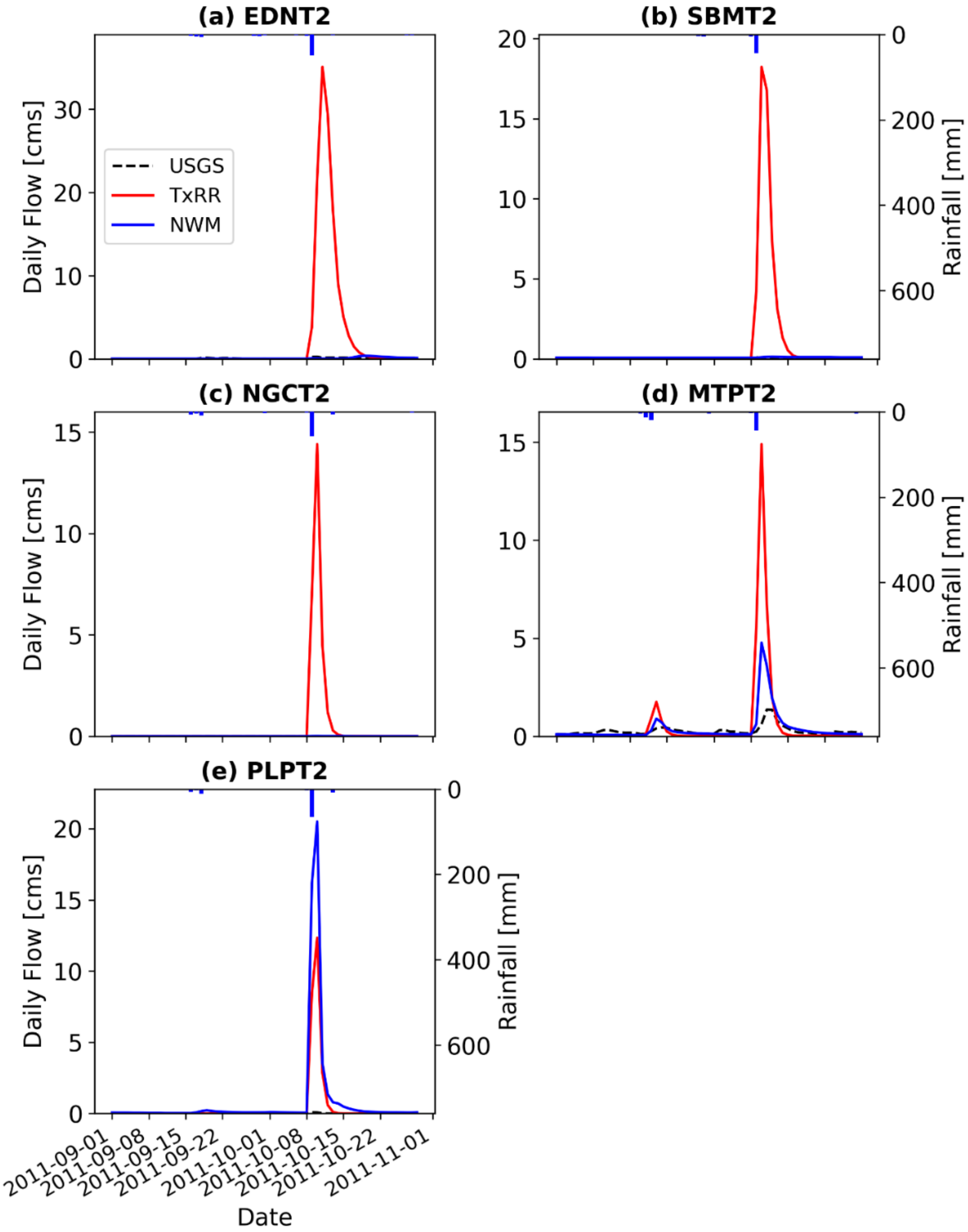


Figure 13: Comparisons of simulated flow produced by TxRR and NWM models and USGS observations for a drought event occurred in September 2011.

3.3. Comparison of parameter transferability

The transferability of model parameters for TxRR and NWM is assessed by examining the model performance for the four Group 2 watersheds.

The overall relative performance of the two models is summarized in Table 4. The observations in some ways mirror those for the Group 1 watersheds. The most striking feature is that, in terms of correlation, TxRR broadly outperforms NWM at daily scale but underperforms the latter at yearly scale. This echoes the finding for Group 1 watershed. Another observation of note is that NWM simulation exhibits severe, negative biases for three out of the four watersheds. For these three watersheds, TxRR simulations are also negatively biased, but with lesser severity. For example, for the watershed GNDT2 (West Mustang Creek), NWM simulation features a bias of -59%, whereas the bias for TxRR simulation is only -19%. The only watershed where the simulations by both models exhibit positive biases is LSNT2 (Navidad River at Strane Park) that is situated in the north. Note that the earlier analysis for Group 1 watersheds shows that simulations from both model exhibit positive biases for SBMT2, also the northmost watershed in that group. What is intriguing, however, is that the SMMAX for LSNT2 was adapted not from SBMT2 but from EDNT2 (Lavaca River), for which TxRR simulation exhibits a severe, negative bias.

Table 4: Validation statistics of daily, monthly, and yearly streamflow for ungauged (i.e., Group-2) watersheds produced by parameter transferred TxRR model and NWM simulations. Dark green, light green, and gold color represent outperformance of the respective model based on Pearson's correlations, RMSE, and percent biases, respectively.

					Daily				
SHEF ID	**R**			**RMSE [cms]**			**%BIAS**		
	NWM	NWM-BC	TxRR	NWM	NWM-BC	TxRR	NWM	NWM-BC	TxRR
LSNT2	0.62	0.62	0.71	25	25	27	24	37	32
CODT2	0.73	0.73	0.83	16	16	12	-23	-23	-13
GNDT2	0.74	0.74	0.83	13	13	10	-59	-56	-19
LMCT2	0.65	0.65	0.83	5	5	4	-46	-35	-9
					Monthly				
SHEF ID	**R**			**RMSE [cms]**			**%BIAS**		
	NWM	NWM-BC	TxRR	NWM	NWM-BC	TxRR	NWM	NWM-BC	TxRR
LSNT2	0.82	0.82	0.80	8	9	12	24	36	32
CODT2	0.76	0.76	0.87	6	6	5	-22	-23	-13
GNDT2	0.81	0.81	0.88	6	6	4	-59	-55	-18
LMCT2	0.83	0.83	0.88	2	2	1	-45	-34	-9
					Yearly				
SHEF ID	**R**			**RMSE [cms]**			**%BIAS**		
	NWM	NWM-BC	TxRR	NWM	NWM-BC	TxRR	NWM	NWM-BC	TxRR
LSNT2	0.95	0.95	0.90	3	4	4	26	39	33
CODT2	0.92	0.92	0.83	2	2	3	-24	-25	-17
GNDT2	0.87	0.87	0.86	3	3	2	-59	-56	-19
LMCT2	0.93	0.93	0.91	1	1	0	-45	-34	-9

Fig. 14 shows the flow duration curves based on the model simulations and observations. It appears that the flow records for all watersheds contain a substantial number of zeros. For LSNT2 and CODT2, the two watersheds in the north, zeros make up about 10% of records, whereas for LMCT2 (East Mustang Creek), the smallest among the watersheds (140 km^2), the stream was dry about 20% of time. For LSNT2 and CODT2, the simulations by both TxRR and NWM tend to inflate the very low flow, though the inflation is more severe for the latter simulation, a feature reminiscent of similar observations for the Group 1 watersheds. For the two smaller watersheds in the south (GNDT2 and LMCT2), NWM accurately reproduces the percentage dry days, whereas TxRR simulations fail to do so. The performance of the two models in resolving the duration of baseflow is mixed. The CVM test again yields mixed results, with NWM outperforming for two watersheds (CODT2 and LMCT2) and underperforming for the other two (LSNT and GNDT2).

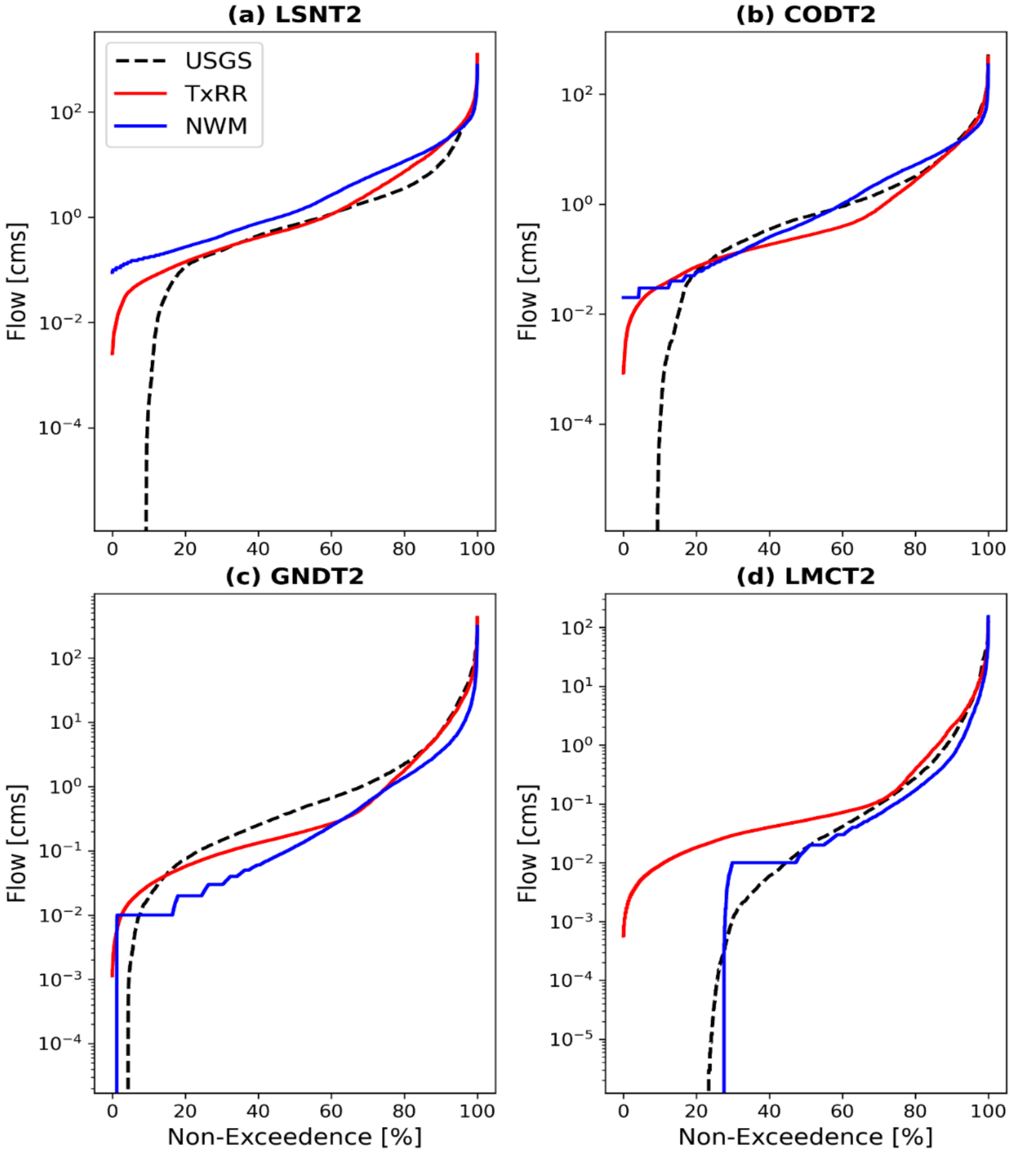


Figure 14: Comparisons of flow duration curve produced by TxRR and NWM models, bias corrected NWM flow and USGS observation for four newly identified stations within the study area.

Taken together, the findings illustrated above suggest that the parameter transfer scheme of TxRR is overall more effective than that for NWM in that TxRR simulations based on the transferred parameter values are a) not as severely biased as the NWM simulations, and b) are closely correlated with observations. Yet, it is apparent that the TxRR has a limitation of not being able to resolve the frequency of dry conditions for intermittent streams. In addition, there is a prominent feature that the TxRR and NWM outperforms at different time scales in terms of correlation – NWM broadly underperforms at daily scale but its performance overtakes that of TxRR at yearly scale, a feature consistent with the observation for Group 1 watersheds.

4. DISCUSSION

A key motivation of the investigation was to determine whether the sophisticated land surface scheme of NWM will confer it the ability to more realistically represent runoff dynamics under extreme weather and climate conditions relative to simpler, conceptual models. While a plethora of efforts have been dedicated to model comparisons (Reed et al., 2004; Smith et al., 2012), studies that directly involve NWM (or WRF-Hydro) have been scarce. Moreover, few of extant works related to NWM investigated its performance for both hydroclimatic extremes. Previous studies that evaluated the performance of NWM for extreme conditions, such as Hansen et al. (2019), focused on the model's ability to capture low flow events. Their findings suggest difficulties of NWM in reproducing these low flow episodes. Specifically, Hansen et al., (2019) underscored a consistent underrepresentation of frequencies of low flow episodes along the Colorado River. By contrast, the present study over the Matagorda Basin near central Texas coast reveals that the NWM in fact performs reasonably well in reproducing the dry conditions during the extreme drought of 2011. In fact, it clearly outperforms the simpler model TxRR by producing little to no runoff following sizable rainfall input through the fall of 2011, whereas the latter vastly oversimulates the runoff over the same time window. We surmise that the relative good performance has to do with a) the ability of NWM in representing the dry soil moisture conditions and abstraction during the drought, and b) the fact that NWM was calibrated for a window that emcompasses the drought episode of 2011-2014 where runoff production was low in the region.

Much is unknown about mechanisms that give rise to the varying performance of NWM illustrated herein and in aforementioned works, though the hydroclimatic conditions during the calibration window may have played a major role. Note that streamflow in the Colorado River Basin is to a large extent modulated by snowmelt rather than rainfall (Gan *et al.*, 2022), and efficacy of calibration over the latter region may be complicated by the mechanistic deficiencies in both snow and water balance models (Gan *et al.*, 2022). By contrast, runoff along the Texas coast is dominated by rainfall and this may have made calibration more effective. In addition, it is worth noting that the differences between the findings from the present study and previous ones may have to do with the differences in the scope of evaluation - the aforementioned works examined low flow events over a much longer time window, whereas the present one focuses on a single episode.

Perhaps the most notable finding of this study is the scale-dependence in the relative performance of the models for the study watersheds. While NWM underperforms TxRR at the daily scale, its performance improves at longer time aggregates and in fact surpasses that of the latter at annual scale. The underpeformance of the NWM at shorter time scales is broadly consistent with the observations from earlier efforts (Reed et al., 2004; Bárdossy, 2007; Coron *et al.*, 2014; Ren *et al.*, 2016), where it was evident that model calibration and parameter transfer both become more challenging with complexity of hydrologic model and the dimension of parameter set. NWM features a sophisticated parameter transfer scheme that relies on established ecoregions and is supplemented by cluster analysis. The robustness of this scheme, however, remains in question, as large biases are apparent in NWM simulations for watersheds near the bay to which the parameters were transferred. By contrast, the transfer scheme of TxRR appears effectual, at least for small regions where differences in runoff potential can be adequately described by those in curve number.

The findings to a large extent corroborate our postulation that NWM offers potentially higher skills in resolving water balance at longer time scales – the nearly uniformly superior correlation of NWM simulations at the yearly scale is a telling sign, and *prima facie* evidence supporting this thesis. In this region, while rainfall variability remains unmistakably the dominant regulator of runoff across time and spatial scales, surface energy balance appears to be a potent modulator of water balance at longer time ranges. This holds true even for some of the southern watersheds underlain by clayish soils with seemingly limited soil water storage capacity. The underperformance of TxRR at longer time scales may be a reflection of its inability to realistically account for the role of soil moisture/temperature in regulating surface latent heat fluxes.

It should be noted that precipitation and surface energy availability are closely intertwined, as weather patterns associated with precipitation systems (cloudiness, near-surface temperature and humidity) and surface conditions (albedo) jointly determine surface radiative energy balance. The detailed mechanisms in play that give rise to the scale-dependent performance differentials await further scrutiny. Yet, it is evident that failures to account for the interannual variations of the energy balance would likely artificially distort the longer-range variability in runoff production.

Another important feature highlighted in the study concerns the inability of both models to reproduce the peak flow during the two major flooding events. The bias in daily flow is particularly negative in NWM reanalysis. Earlier studies assessing NWM analysis also reported similar negative biases. Johnson *et al.*, (2019), for example, found that the inundation maps generated by combining NWM 1.2 reanalysis with the HAND method (Nobre *et al.*, 2016) underpresernted the extent of inundation, and cited the underprediction of streamflow by NWM as one of the primary causes. It should be noted, however, that this underprediction can be at least partially explained by the negative biasse in NLDAS-2 precipitation data during these events - there is a negative bias in the NLDAS precipitation data evident in major storm events that is indicative of data quality issues. Thus far, it is not clear whether model physics or calibration strategies have played roles in producing the bias in NWM reanalysis as well as TxRR simulation, and further investigations that use more accurate precipitation products, e.g., the NWS multisensor product (Zhang *et al.*, 2011), are warranted.

5. CONCLUSIONS

This study assesses the relative accuracy of two hydrologic models, namely the Texas Rainfall-Runoff (TxRR) and the National Water Model (NWM) in reproducing historical flow over a group of watersheds near central Texas coast. The study region is well-known for large swings in hydroclimatic conditions and experienced some of the extreme flooding and drought conditions over the period of 1993-2017. The relative performance of the two models was characterized by aggregate statistics and for two major flooding events and an extreme drought episode in modern history.

Our basic hypothesis was that NWM's sophisticated representation of surface energy balance and soil water dynamics would be advantageous would allow it to better capture the runoff dynamics than the simpler, conceptual TxRR. Our experiments yielded mixed results. At shorter (daily and monthly) scales, TxRR outperforms NWM in resolving the runoff dynamics, while the opposite is true at the annual scale. This distinct temporal scale dependence in the relative performance of TxRR and NWM is rather unexpected, and to the best of knowledge of the authors, has not been reported or systematically examined. The outperformance of TxRR at shorter time scales is likely reflective of its superior calibration which compensats for its structural simplicity and masks the model's underrepresentation of physical processes. This advantage in calibration, however, diminishes at interannual time scale where NWM clearly outperforms across a majority of watersheds, highlighting the merit of its explicity characterization of surface energy balance that regulates soilwater dynamics at longer temporal scales. Specific processes that give rise to this scale-dependent performance warrant future investigations.

The study mostly confirms the trade-off between model complexity and robustness of calibration/parameter transfer illustrated in Reed et al. (2004), and Zhang and Shuster (2014). It is evident that model complexity may degrade the efficacy of parameter transfer scheme. This finding has broad implications for estimation of freshwater inflow for coastal watersheds where observational data are scarce, and for which transferability of parameters is a key determinant of the accuracy of estimates. While the accuracy of NWM retrospective analysis has been gradually improving thanks partially to expanded calibration efforts, much remains to be done to enhance the present mechanism for regionalizing watersheds and transferring parameter values. In particular, a more in-depth analysis will be helpful to diagnose the transfer scheme among watersheds with contrasting contributions of infiltration excess runoff that most likely is a key player over watersheds with significant presence of clayish soil.

The case studies point to contrasting performance of models over flooding and drought conditions. Both models were unable to reproduce the peak daily discharge over both events for the five watersheds in Group 1. This underestimation is in part stems from severely negative biases in the NLDAS-2 precipitation product that are indicative of quality control problems – the authors surmise that missing gauge data may have been a cause of the problem.   On the other hands, the way the models were calibrated may have compounded the negative biases, as it is known that calibration using metrics such as RMSE tends to introduce or exacerbate the conditional bias (Seo, 2013). Between the models, there appears to be a consistent tendency for NWM to underpredict the rising limb of the hydrographs, hinting a possibility that NWM produced consistently drier antecedent soil moisture conditions that underpinned its underprediction of earlier hydrograph rises. Future experiments, in which alternative precipitation data sets that exhibit lesser biases seve as forcings to drive both models, will help discern potential issues in model structures or calibration approaches, and determine their relative effectiveness in capturing high flow events.   Note that the biases in the NLDAS precipitation data most likely were far less severe during the 2011 drought as the magnitude of rainfall was low.   Examining soil moisture against remotely sensing product relative to soil moisture contents represented in each model will help shed light on the model realisms in representing the evapotranspiration during these events, and can guide future model calibration and structural improvements.

Finally, the scope of the study did not permit a close analysis of the impacts of subdaily rainfall structure on TxRR model errors, or the roles in differential contributions of interflow and groundwater discharge in runoff production.   These will be left to future efforts.

ACKNOWLEDGEMENTS

This work is supported by the Texas Water Development Board contract #1800012276.   The first author was partially supported by the faculty startup package of Dr. Yu Zhang at UT Arlington, and by NOAA Grant NA19OAR4310347.   The authors are thankful to Brian Cosgrove and Yuqiong Liu at the NWS Office of Water Prediction, and Kris Lander at NWS West Gulf River Forecast Center for providing information on the NWM model calibration. Ram Neupane at TWDB assisted with acquisition of the TWDB mean areal precipitation data set and this effort is acknowledged here.

REFERENCES

Bárdossy, A., 2007. Calibration of Hydrological Model Parameters for Ungauged Catchments. Hydrology and Earth System Sciences 11. doi:10.5194/hess-11-703-2007.

Beven, K.J. and M.J. Kirkby, 1979. A Physically Based, Variable Contributing Area Model of Basin Hydrology. Hydrological Sciences Bulletin 24:43–69.

Budyko, M. I., D.H. Miller, and D.H. Miller, 1974. Climate and Life Budyko (Editor). New York: Academic Press.

Budyko, M.I., 1961. The Heat Balance of the Earth's Surface. Soviet Geography 2. doi:10.1080/00385417.1961.10770761.

Burnash, R.J.C., R.L. Ferral, and R.A. McGuire, 1973. A Generalized Streamflow Simulation System – Conceptual Modeling for Digital Computers. U.S. Dept. of Commerce National Weather Service and State of California Department of Water Resources:204.

Chen, F. and J. Dudhia, 2001. Coupling and Advanced Land Surface-Hydrology Model with the Penn State-NCAR MM5 Modeling System. Part I: Model Implementation and Sensitivity. Monthly Weather Review 129:569–585.

Coron, L., V. Andréassian, C. Perrin, M. Bourqui, and F. Hendrickx, 2014. On the Lack of Robustness of Hydrologic Models Regarding Water Balance Simulation: A Diagnostic Approach Applied to Three Models of Increasing Complexity on 20 Mountainous Catchments. Hydrology and Earth System Sciences 18. doi:10.5194/hess-18-727-2014.

Cosgrove, B., Gochis D., E.P. Clark, Z. Cui, A.L. Dugger, X. Feng, L.R. Karsten, S. Khan, D. Kitzmiller, H.S. Lee, and Y. Liu, 2016. An Overview of the National Weather Service National Water Model.

Cronshey, R.G., R.T. Roberts, and N. Miller, 1985. Urban Hydrology For Small Watersheds (TR-55 REV.). In Hydraulics and Hydrology in the Small Computer Age. ASCE:1268–1273.

Crow, W.T., F. Chen, R.H. Reichle, Y. Xia, and Q. Liu, 2018. Exploiting Soil Moisture, Precipitation, and Streamflow Observations to Evaluate Soil Moisture/Runoff Coupling in Land Surface Models. Geophysical Research Letters 45. doi:10.1029/2018GL077193.

Daly, C., R.P. Neilson, and D.L. Phillips, 1994. A Statistical-Topographic Model for Mapping Climatological Precipitation over Mountainous Terrain. Journal of Applied Meteorology 33:140–158.

Duan, Q., J. Schaake, V. Andréassian, S. Franks, G. Goteti, H. V. Gupta, Y.M. Gusev, F. Habets, A. Hall, L. Hay, T. Hogue, M. Huang, G. Leavesley, X. Liang, O.N. Nasonova, J. Noilhan, L. Oudin, S. Sorooshian, T. Wagener, and E.F. Wood, 2006. Model Parameter Estimation Experiment (MOPEX): An Overview of Science Strategy and Major Results from the Second and Third Workshops. Journal of Hydrology 320:3–17.

Ek, M.B., K.E. Mitchell, Y. Lin, E. Rogers, P. Grunmann, V. Koren, G. Gayno, and J.D. Tarpley, 2003. Implementation of Noah Land Surface Model Advances in the National Centers for Environmental Prediction Operational Mesoscale Eta Model. Journal of Geophysical Research: Atmospheres 108. doi:10.1029/2002jd003296.

Gan, Y., X.Z. Liang, Q. Duan, F. Chen, J. Li, and Y. Zhang, 2019. Assessment and Reduction of the Physical Parameterization Uncertainty for Noah-MP Land Surface Model. Water Resources Research 55:5518–5538.

Gan, Y., Y. Zhang, Y. Liu, C. Kongoli, and C. Grassotti, 2022. Assimilation of Blended in Situ-Satellite Snow Water Equivalent into the National Water Model for Improving Hydrologic Simulation in Two US River Basins. Science of the Total Environment 838:156567.

Gochis, D.J., M. Barlage, R. Cabell, M. Casali, A. Dugger, K. FitzGerald, M. McAllister, J. McCreight, A. RafieeiNasab, L. Read, K. Sampson, D. Yates, and Y. Zhang, 2020. The WRF-Hydro Modeling System Technical Description, (Version 5.1.1). NCAR Technical Note.

Gochis, D.J., W. Yu, and D. Yates, 2015. The WRF-Hydro Model Technical Description and User's Guide, Version 3.0. NCAR Technical Document.

Hansen, C., S.J. Shafiei, S. McDonald, and A. Nabors, 2019. Assessing Retrospective National Water Model Streamflow with Respect to Droughts and Low Flows in the Colorado River Basin. Journal of the American Water Resources Association 55. doi:10.1111/1752-1688.12784.

Hawkins, R.H., 1978. Runoff Curve Numbers With Varying Site Moisture. ASCE J Irrig Drain Div 104. doi:10.1061/jrcea4.0001221.

Johnson, J.M., D. Munasinghe, D. Eyelade, and S. Cohen, 2019. A Comprehensive Evaluation of the National Water Model (NWM) – Height Above Nearest Drainage (HAND) Flood Mapping Methodology. Natural Hazards and Earth System Sciences:1–17.

Kim, H.C. and P.A. Montagna, 2009. Implications of Colorado River (Texas, USA) Freshwater Inflow to Benthic Ecosystem Dynamics: A Modeling Study. Estuarine, Coastal and Shelf Science 83:491–504.

Kohler, M.. and R.. Linsley, 1951. Predicting the Runoff from Storm Rainfall. Weather Bureau Research Paper No. 34:10.

Kuzmin, V., D.J. Seo, and V. Koren, 2008. Fast and Efficient Optimization of Hydrologic Model Parameters Using a Priori Estimates and Stepwise Line Search. Journal of Hydrology 353:109–128.

Manabe, S., 1969. Climate and The Ocean Circulation 1: I. The Atmospheric Circulation and The Hydrology of The Earth's Surface. Monthly Weather Review 97:739–774.

Matsumoto J., 1992. User's Manual for the TWDB's Rainfall-runoff Model. AGU. https://www.twdb.texas.gov/. Accessed 13 Mar 2021.

Michael Johnson, J., D. Munasinghe, D. Eyelade, and S. Cohen, 2019. An Integrated Evaluation of the National Water Model (NWM)-Height above Nearest Drainage (HAND) Flood Mapping Methodology. Natural Hazards and Earth System Sciences 19. doi:10.5194/nhess-19-2405-2019.

Mishra, S.K. and V.P. Singh, 2004. Validity and Extension of the SCS-CN Method for Computing Infiltration and Rainfall-Excess Rates. Hydrological Processes 18:3323–3345.

Nasab, A.R., L. Karsten, A. Dugger, K.F.R. Cabell, D. Gochis, D. Yates, K. Sampson, J. McCreight, L. Read, Y. Zhang, and M. McAllister, 2020. Overview of National Water Model CalibrationGeneral Strategy & Optimization. NCAR. https://ral.ucar.edu/sites/default/files/public/projects/wrf-hydro/training-materials/calibrationnov2020-arezoo.pdf.

Nash, J.E. and J. V. Sutcliffe, 1970. River Flow Forecasting through Conceptual Models Part I - A Discussion of Principles. Journal of Hydrology 10:282–290.

Niu, G.Y., Z.L. Yang, K.E. Mitchell, F. Chen, M.B. Ek, M. Barlage, A. Kumar, K. Manning, D. Niyogi, E. Rosero, M. Tewari, and Y. Xia, 2011. The Community Noah Land Surface Model with Multiparameterization Options (Noah-MP): 1. Model Description and Evaluation with Local-Scale Measurements. Journal of Geophysical Research Atmospheres 116:1–19.

Nobre, A.D., L.A. Cuartas, M.R. Momo, D.L. Severo, A. Pinheiro, and C.A. Nobre, 2016. HAND Contour: A New Proxy Predictor of Inundation Extent. Hydrological Processes 30. doi:10.1002/hyp.10581.

Ogden, F.L. and P.Y. Julien, 1994. Runoff Model Sensitivity to Radar Rainfall Resolution. Journal of Hydrology 158:1–18.

Reed, S., V. Koren, M. Smith, Z. Zhang, F. Moreda, and D.J. Seo, 2004. Overall Distributed Model Intercomparison Project Results. Journal of Hydrology. doi:10.1016/j.jhydrol.2004.03.031.

Ren, H., Z. Hou, M. Huang, J. Bao, Y. Sun, T. Tesfa, and L. Ruby Leung, 2016. Classification of Hydrological Parameter Sensitivity and Evaluation of Parameter Transferability across 431 US MOPEX Basins. Journal of Hydrology 536. doi:10.1016/j.jhydrol.2016.02.042.

Russell, M.J., P.A. Montagna, and R.D. Kalke, 2006. The Effect of Freshwater Inflow on Net Ecosystem Metabolism in Lavaca Bay, Texas. Estuarine, Coastal and Shelf Science 68:231–244.

Schaake, J.C., V.I. Koren, Q.Y. Duan, K. Mitchell, and F. Chen, 1996. Simple Water Balance Model for Estimating Runoff at Different Spatial and Temporal Scales. Journal of Geophysical Research Atmospheres 101:7461–7475.

Scharffenberg, W., 2016. Hydrologic Modeling System HEC-HMS User's Manual. U.S. Army Corps of Engineers - Hydrologic Engineering Center 1.

Schoenbaechler C., Guthrie C. G., Matsumoto J., Lu Q., and Negusse S., 2011. Model Calibration and Validation for the Lavaca-Colorado Estuary and East Matagorda Bay. Austin.

Seo, D.J., 2013. Conditional Bias-Penalized Kriging (CBPK). Stochastic Environmental Research and Risk Assessment 27. doi:10.1007/s00477-012-0567-z.

Smith, M.B., V. Koren, S. Reed, Z. Zhang, Y. Zhang, F. Moreda, Z. Cui, N. Mizukami, E.A. Anderson, and B.A. Cosgrove, 2012. The Distributed Model Intercomparison Project - Phase 2: Motivation and Design of the Oklahoma Experiments. Journal of Hydrology 418–419:3–16.

Tolson, B.A. and C.A. Shoemaker, 2007. Dynamically Dimensioned Search Algorithm for Computationally Efficient Watershed Model Calibration. Water Resources Research 43. doi:10.1029/2005WR004723.

Viterbo, F., K. Mahoney, L. Read, F. Salas, B. Bates, J. Elliott, B. Cosgrove, A. Dugger, D. Gochis, and R. Cifelli, 2020a. A Multiscale, Hydrometeorological Forecast Evaluation of National Water Model Forecasts of the May 2018 Ellicott City, Maryland, Flood. Journal of Hydrometeorology 21:475–499.

Viterbo, F., L. Read, K. Nowak, A.W. Wood, D. Gochis, R. Cifelli, and M. Hughes, 2020b. General Assessment of the Operational Utility of National Water Model Reservoir Inflows for the Bureau of Reclamation Facilities. Water (Switzerland) 12:1–30.

Wickham, J., C. Homer, J. Vogelmann, A. McKerrow, R. Mueller, N. Herold, and J. Coulston, 2014. The Multi-Resolution Land Characteristics (MRLC) Consortium - 20 Years of Development and Integration of USA National Land Cover Data. Remote Sensing 6:7424–7441.

Williams, J.R. and W. V. LaSeur, 1976. Water Yield Model Using SCS Curve Numbers. ASCE J Hydraul Div 102. doi:10.1061/jyceaj.0004609.

Xia, Y., K. Mitchell, M. Ek, J. Sheffield, B. Cosgrove, E. Wood, L. Luo, C. Alonge, H. Wei, J. Meng, B. Livneh, D. Lettenmaier, V. Koren, Q. Duan, K. Mo, Y. Fan, and D. Mocko, 2012. Continental-Scale Water and Energy Flux Analysis and Validation for the North American Land Data Assimilation System Project Phase 2 (NLDAS-2): 1. Intercomparison and Application of Model Products. Journal of Geophysical Research Atmospheres 117. doi:10.1029/2011JD016048.

Yang, Z.L., G.Y. Niu, K.E. Mitchell, F. Chen, M.B. Ek, M. Barlage, L. Longuevergne, K. Manning, D. Niyogi, M. Tewari, and Y. Xia, 2011. The Community Noah Land Surface Model with Multiparameterization Options (Noah-MP): 2. Evaluation over Global River Basins. Journal of Geophysical Research Atmospheres 116:1–16.

Zhang, Y., S. Reed, and D. Kitzmiller, 2011. Effects of Retrospective Gauge-Based Readjustment of Multisensor Precipitation Estimates on Hydrologic Simulations. Journal of Hydrometeorology 12:429–443.

Zhang, Y. and W. Shuster, 2014. The Comparative Accuracy of Two Hydrologic Models in Simulating Warm-Season Runoff for Two Small, Hillslope Catchments. Journal of the American Water Resources Association 50:434–447.