Dear Editor Kurtz and Reviewer #2,

We would like to thank you again for the helpful comments and suggestions. In this version of revised manuscript, we did following changes to address the comments and suggestions:

1. We adjust the observation of streamflow for the model domain, and re-do the WRF-Hydro calibration for the same study area and same stations. Tables and figures are all updated accordingly.
2. Alternatively, we also calibrate WRF-Hydro over the same study area but using local river stations that are included in the study area. Tables and figures are presented in Supporting Information.
3. We present scalability analysis for the computational benefits of the HPC-enabled PEST coupled with WRF-hydro. A new figure is added. Table 3 is extended to include more scenarios.
4. Detailed comments are added in the published code to advise the use of developed scripts on other management software or job schedulers.

Please find our one-on-one response below to both editor's and reviewer's comment. A complete list of the changes made for the revised manuscript can be found in the "track changes" version of the manuscript. A clean version of the revised manuscript is also attached at the end.

Sincerely, Jiali
Wang
jialiwang@anl.gov

From Editor Kurtz:

(1) A more detailed analysis of the scaling behaviour and the optimal configuration with respect to workers and nodes per WRF-Hydro instance should be performed.

From Reviewer #2:

Concerning the analysis of computational benefits of parallel PEST on HPCs (new Section 4.2), the analysis performed is interesting, but to me it should be still improved, because does not yet respond clearly to the main question: "How can I use the nodes I have available in the most efficient way?" If I understand this section, based on Table 3 the Test 5 is better than the Test 1 because: Test 1) 23 workers x 2 nodes = 46 nodes and 103 min; Test 5) 6 workers x 6 nodes = 36 nodes and only 86 min. On the one hand, this result should be highlighted. But, on the other hand, the same result cannot be separated from considering the extent of the WRF-Hydro domain (from this point of view, I note that information about computational domain is still missing). Therefore, in general I would say that: first, a scalability analysis of WRF-Hydro over the specific computational domain should be performed; then, this scalability analysis should be used as a preliminary, but essential piece of information to provide comprehensive indications about the "optimal" configuration of the system for the analysed case study, given the threshold of the computational resources available.

Response: Thank you both for the great suggestion, which really adds value on our existing analysis about the computational benefits of the HPC-enabled PEST coupled with WRF-Hydro. The domain size had been mentioned a couple of times in Section 2.1 and 2.2. We emphasize it again in this version:

"The domain size is ~495,000 km2 (747 km from west to east; and 657 km from south to north)."

"…hydrological routing is performed at a grid resolution of 200 m, with 3285 south-north × 3735 west-east grid cells"

We add following discussion in Section 4.2 (Computational benefits of parallel PEST on HPCs) and a new figure (Figure 5) in the revised manuscript to demonstrate the scalability analysis. We also did one more experiment using 6 workers and 8 nodes per worker. Based on all the tests we did (Tests 1-6), we provide more scenarios about their time and computing cost when using different number of workers and nodes per worker by extrapolating the existing numbers (Table 3).

"While these numbers in Table 3 and Figure 4 are helpful to demonstrate the scale-up capability of each component (PEST and WRF-Hydro), they do not answer questions such as, if one has certain number of nodes, how many workers and how many nodes per worker should be used to achieve the highest efficiency of the WRF-Hydro calibration using HPC-enabled PEST? On the other hand, one may have unlimited computational resource, but would like to complete the calibration in a short time period. We present scalability analysis below to answer these questions. First, we generate more scenarios using different number of workers and nodes per worker by extrapolating the existing time and computing costs based on the experiments that are already conducted. These scenarios use 23 or 12 workers, and 4, 6, or 8 nodes per worker, respectively. Since we have conducted simulations using the same number of nodes per worker, the cost for these scenarios are easily predicted.

As shown in Figure 5, compare with Test 3 (which requires the least computing resource —12 nodes in total), having more workers (with the same number of nodes for each worker, e.g., Tests 1 and 2), takes more time than the ideal curve. The ideal curve assumes a linear speedup based on the time cost of Test 3. However, using the same number of workers and increasing the number of nodes for each worker (e.g., Tests 4, 5, and 6) can achieve the ideal speedup. Even when using 12 workers, increasing the number of nodes for each worker can still achieve a speedup close to the ideal curve. Using 23 workers will not achieve the ideal speedup. Therefore, if one only has a certain number of nodes available, we recommend to use relatively small number of workers but large number of nodes for each worker. For example, if one has 48 nodes, then there are three options can be considered: using 23 workers and 2 nodes per worker; 12 workers and 4 nodes per worker, and 6 workers and 8 nodes per worker. Other partition (16x3; or 8x6) between numbers of workers and nodes per worker are not as efficient as above. These three options will cost 103, 72 and 60 min, respectively, to finish one iteration. Thus, using 6 workers and 8 nodes per worker is the most efficient way to consume the limited computing resource. On the other hand, if one would like to conduct the calibration in a short time period without any limits for the computing resource, then using 23 workers and 8 nodes (perhaps even more nodes depending on the size of the model domain and the scale up capability of WRF-Hydro), will finish one iteration in ~24 min."

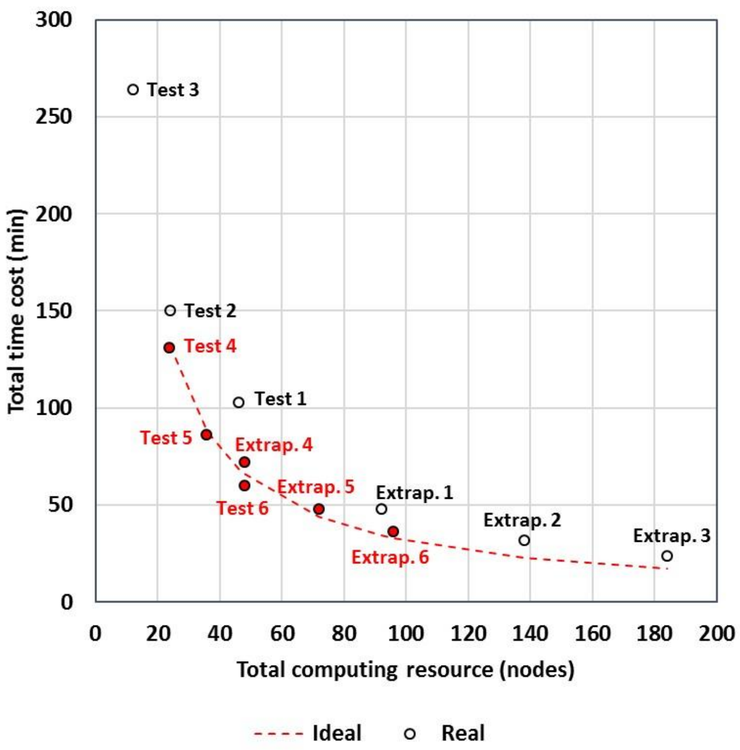

Figure 5. Total time cost and total computing resource needed for each test and extrapolated scenario, which uses different number of workers and different number of nodes per worker. The dash line is an ideal curve, which assumes a liner decrease in terms of time cost when more computing resource is used, built on Test 3. The circles are real cost for time and computing resource by each test and extrapolated. The red text and solid circles indicate those specific tests meet the ideal expectation of speedup.

From Editor Kurtz:

(2) The adopted calibration procedure as well as the description of the catchment models needs to be improved. Please see the comments from reviewer #2 for more details.

From Reviewer #2:

Concerning my doubts about the calibration procedure, the explanation provided: “The reason is that the water contributions for these stations are from a larger river basin (Mississippi River) than we included in our current study area” confirms that the calibration performed over stations from 2 to 4 does not make too much sense. Of course enlarging the study area to include the MRB will improve results, and this must be done if the authors still want to consider stations from 2 to 4. Another option, maybe more feasible for the authors, is to skip stations from 2 to 4 and consider stations from 5 to 8 not only for “transferability”, but also for validation (but this must be done in a persuasive way). However, since from the first version of the manuscript it was not clear that the study area “covers only half of the MRB”, and in the second version the authors only hint at it, it is very useful, for the sake of clarity, that the characteristics of all the catchments upstream the considered stations are clearly showed in the paper, both in a Table (e.g., indicating extent of the catchment and other main features) and in Figure 1, highlighting the borders of the catchments.

Response: Thanks for your comment, which motivates us to think about addressing this issue in two different approaches. First approach: as we presented in the revised manuscript, we adjusted the observation of streamflow for Stations 2, 3, and 4 by excluding inflows from catchments that are not covered by the study domain; Second approach: we also mentioned it in the revised manuscript, but provided results in Supporting Information. We calibrate the model against local river stations which has smaller drainage area and are included in the study area.

Below are the description that we include in the revised manuscript. <u>This is for the first approach</u>. We also edited the Figure 1 to demonstrate the idea, and highlight the borders of the catchments that we work with.

"As shown by the lower left index map in Figure 1, the study area (the red box) only covers the lower part of Upper Mississippi River Basin (UMRB) and a portion of Missouri River Basin (MORB). In order to prepare observation datasets of streamflow contributed only from the drainage area within the model domain, we identified inflows entering the model domain at three different sites, namely, sites 05411500, 06807000, and 06887500, as indicated by the black solid triangles in the index map of Figure 1. The outflows of combined UMRB and MORB can be found at the three outlets, namely, sites 07010000, 07020500, and 07022000 (named Stations 2, 3, and 4, respectively, as shown by black solid circles in Figure 1). These outlets are located sequentially at the main Mississippi River after confluence of Mississippi River and Missouri River. Thus, the observed streamflow contributed by drainage area within the model domain can be calculated by subtracting the sum of the discharge at the three sites (black triangles; recognized as inflow) from the discharge at each of the three outlet sites (black circles; recognized as outflow). The final derived observations of streamflow (or adjusted streamflow observation data) from the drainage area within this model domain are prepared for model calibration and validation. To prove this concept, we validated the consistency of the sum of observed drainage areas at inflow sites plus modeled drainage area with the overall drainage area at the outlet. The drainage area (UMRB and MORB) at outlet site 07010000 is 1.8E+12 m2. The sum of drainage areas at three inflow sites is about 1.4E+12 m2 (2.0E+11, 1.1E+12, and 1.4E+11 m2 for site 05411500, 06807000, and 06887500, respectively) and the modeled drainage area is 0.36E+12 m2; the total area is 1.76+12 m2. This indicates that the flows from sum of three inflow sites and modeled result represent 98% of drainage area at the outflow site 07010000. Therefore, the adjusted streamflow observation data are qualified for model calibration."

<u>For the second approach:</u>

"Alternatively, instead of calibrating the stations that have large drainage area and water coming from outside of the current model domain, we have also tested calibrating small flows at local stations that have relatively small drainage area covered by the current study area. This requires to generate a new high-resolution GIS data file to distribute the stations of interest. We first run the WRF-Hydro model for 6 month using default parameters to spin up the model, and then we calibrate the model based on observations of these local stations. Results including figures and tables are shown in Supporting Information. The calibration results are improved compared to the results that use default parameters, although further improvements are still needed. This again may be because the parameter range are not wide enough to consider the possible values of parameters that work for these specific areas represented at local stations, as we see many optimal parameters hit the bound of the parameter range. More tests to figure out a better set of parameters are needed for future investigation, which is beyond the scope of this study, since our goal is to present the feasibility of HPC enabled PEST."

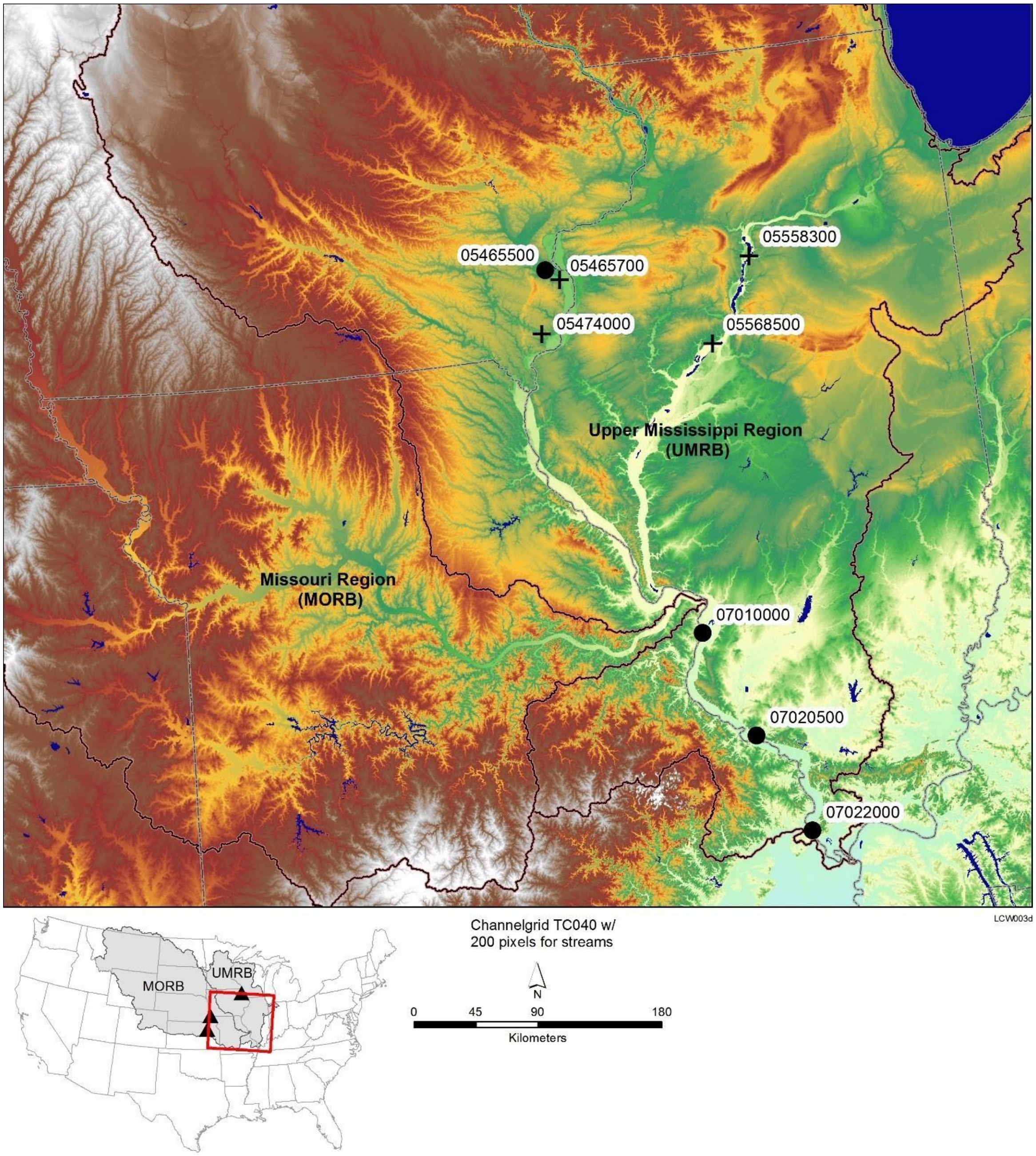

Figure 1: Eight USGS sites over the study area (750 km x 660 km). The four circles are sites that are used for calibrations; the four crosses are sites that are used for transferability assessment. USGS site numbers corresponding to the site index used in this study are: Station 1: 05465500; Station 2: 07010000; Station 3: 07020500; Station 4: 07022000; Station 5: 05465700; Station 6: 05474000; Station 7: 05558300; Station 8: 05568500. The three inflow stations indicated by the black solid triangles are 06807000, 06887500, and 05389500.

Table S3: Statistics of model performance using optimum and default (in parentheses) parameters for local stations during the calibration period.

Statistics	0556855	05474000	05465500	05558300
		Calibration		
NSE	**0.30 (-0.46)**	**-8.5 (-46.6)**	**0.45 (-1.28)**	**-7.11 (-16.6)**
RMSE	**431.88 (624.95)**	**235.25 (526.44)**	**351.18 (716.0)**	**1579.9 (2329.9)**
PCC	**0.96 (0.30)**	**0.54 (-0.86)**	**0.78 (0.33)**	**0.70 (-0.23)**

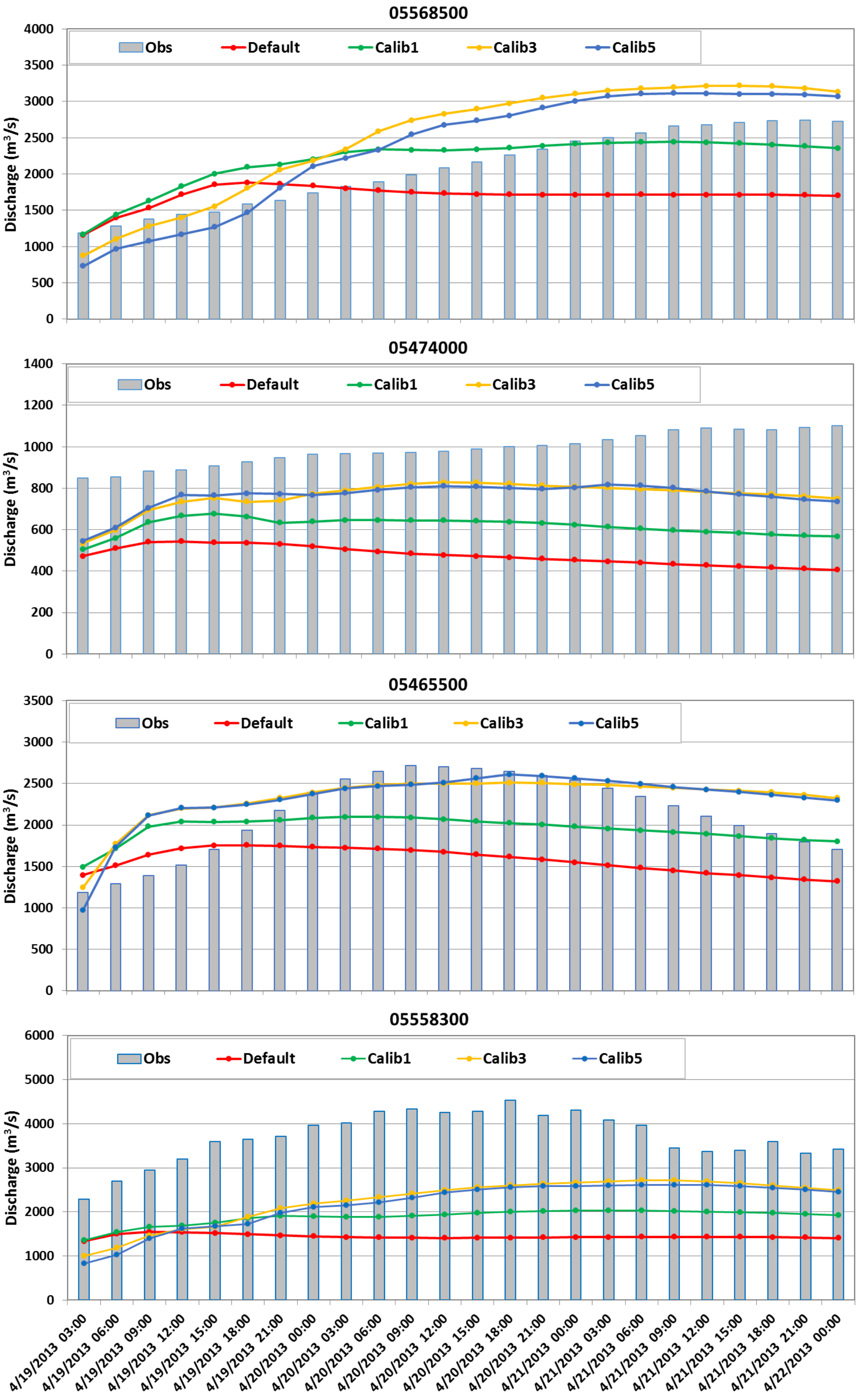
05568500
Obs
Default
Calib1
Calib3
Calib5
Discharge (m³/s)
4000
3500
3000
2500
2000
1500
1000
500
0
05474000
Obs
Default
Calib1
Calib3
Calib5
Discharge (m³/s)
1400
1200
1000
800
600
400
200
0
05465500
Obs
Default
Calib1
Calib3
Calib5
Discharge (m³/s)
3500
3000
2500
2000
1500
1000
500
0
05558300
Obs
Default
Calib1
Calib3
Calib5
Discharge (m³/s)
6000
5000
4000
3000
2000
1000
0
4/19/2013 03:00
4/19/2013 06:00
4/19/2013 09:00
4/19/2013 12:00
4/19/2013 15:00
4/19/2013 18:00
4/19/2013 21:00
4/20/2013 00:00
4/20/2013 03:00
4/20/2013 06:00
4/20/2013 09:00
4/20/2013 12:00
4/20/2013 15:00
4/20/2013 18:00
4/20/2013 21:00
4/21/2013 00:00
4/21/2013 03:00
4/21/2013 06:00
4/21/2013 09:00
4/21/2013 12:00
4/21/2013 15:00
4/21/2013 18:00
4/21/2013 21:00
4/22/2013 00:00

Figure S2: Observed and modeled discharge (m^3/sec) using default and calibrated parameters during a 3-day calibration period (April 19–21, 2013) over four local stations. Station numbers are indicated on top of each panel.

From Editor Kurtz:

(3) I would also like to encourage you to provide a bit more details on how the interface between WRF-Hydro and PEST works on a technical level (i.e., add a bit more description on what happens in the developed script) and maybe add a short discussion on what would need to be done to use the interface on other systems than the ones described in the paper. This would help potential users to adopt the described interface to their own system.

Response: Thank you for the suggestion. The script we developed to bridge the parallel PEST to WRF-hydro on HPC is fairly straightforward and easy to use. While we had described what the script does, in this version, we emphasize that, if it is going to be used on a different management software using a different job scheduler, the only two things that users need to figure out are: (1) how to find and identify available nodes, and (2) how to submit a regular job on that specific server. Although the experiments are presented are conducted on SLURM manager and job scheduler, we have also provided scripts that have been tested and work well on Cobalt manager and job scheduler, which provide the differences in scripts that operate on different HPCs. We have also added detailed comments in the published code (http://doi.org/10.5281/zenodo.3247116) which advise users to make the changes if needed. We are also open to be contacted if users need help from our side. We look forward to working with WRF-Hydro users to make this tool feasible on different HPCs and for different user cases.

We would like to thank the reviewer and the editor again for your suggestions and comments, which tremendously improve this study.

A parallel workflow implementation for PEST version 13.6 in high-performance computing for WRF-Hydro version 5.0: a case study over the midwestern United States

[1]Jiali Wang, [1]Cheng Wang, [2]Vishwas Rao, [1]Andrew Orr, [1]Eugene Yan, [1]Rao Kotamarthi

[1]Argonne National Laboratory, Environmental Science Division, 9700 South Cass Avenue, Lemont, IL 60439, USA

[2]Argonne National Laboratory, Mathematics and Computer Science Division, 9700 South Cass Avenue, Lemont, IL 60439, USA

Correspondence to: Jiali Wang ~~(jialiwang@anl.gov)~~(jialiwang@anl.gov); Rao Kotamarthi (vrkotamarthi@anl.gov)

Abstract. The Weather Research and Forecasting Hydrological (WRF-Hydro) system is a state-of-the-art numerical model that models the entire hydrological cycle based on physical principles. As with other hydrological models, WRF-Hydro parameterizes many physical processes. Hence, WRF-Hydro needs to be calibrated to optimize its output with respect to observations for the application region. When applied to a relatively large domain, both WRF-Hydro simulations and calibrations require intensive computing resources and are best performed on multimode, multicore high-performance computing (HPC) systems. Typically, each physics-based model requires a calibration process that works specifically with that model and is not transferrable to a different process or model. The parameter estimation tool (PEST) is a flexible and generic calibration tool that can be used in principle to calibrate any of these models. In its existing configuration, however, PEST is not designed to work on the current generation of massively parallel HPC clusters. To address this issue, we ported the parallel PEST to HPCs and adapted it to work with WRF-Hydro. The porting involved writing scripts to modify the workflow for different workload managers and job schedulers, as well as developing code to connect parallel PEST to WRF-Hydro. To test the operational feasibility and the ~~potential~~ computational benefits of this first-of-its-kind HPC-enabled parallel PEST, we developed a case study using a flood in the midwestern United States in 2013. Results on a problem involving calibration of 22 parameters show that on the same computing resource used for parallel WRF-Hydro, the HPC-enabled parallel PEST can speed the calibration process by a factor of up to 15 compared with commonly used

PEST in sequential mode. The speedup factor is expected to be greater with a larger calibration problem (e.g., more parameters to be calibrated or a larger size of study area).

1 Introduction

Physically based hydrological models contain detailed physical mechanisms to model the hydrological cycle, but many complex physical processes in these models are parameterized. For example, the state-of-the-art Weather Research and Forecasting Hydrological (WRF-Hydro) modeling system (Gochis et al., ~~2015~~2018) has dozens of parameters that can be land- and river-type dependent and are typically specified in lookup tables. Therefore, these hydrological models need to be calibrated before they can be applied to research over different regions. In this context, calibration refers to adjusting the values of the model parameters so that the model can closely match the behavior of the real system it represents. In some cases, the appropriate value for a model parameter can be determined through direct measurements conducted on the real system. In many situations, however, the model parameters are conceptual representations of abstract watershed characteristics and must be determined through calibration. In fact, model calibration is the most time-consuming step, not only for hydrological models, but also for Earth system model development, because both parametric estimation and parametric uncertainty analysis require hundreds—if not thousands—of model simulations to understand how perturbations in model parameters affect simulations of dominant physical processes and to find the optimum value of a single parameter.

WRF-Hydro is a numerical model that can simulate the entire hydrological cycle using advanced high-resolution data such as satellite and radar products. Compared with the traditional land surface model (LSM) used by WRF, WRF-Hydro provides a framework for multiscale representation of surface flow, subsurface flow, channel routing, and baseflow, as well as a simple lake/reservoir routing scheme. As a physics-based model, WRF-Hydro includes many complicated physical processes that are nonlinear and must be parameterized. The default parameters given by WRF-Hydro may be valid for one region but not for another region. Hence calibration of related model parameters is often required in order to use the model in a new domain. In particular, for a large spatial domain such as the entire contiguous United States, in order to develop the optimal parameter sets in a reasonable amount of time, the calibration must be conducted on high-performance computing (HPC) systems in parallel instead of in the traditional sequential mode. To date, no such calibration tool can efficiently calibrate WRF-Hydro on HPC resources. Typically, each physics-based model needs a calibration code that is custom designed to work with that particular numerical model and its set of physics parameterizations, software architecture, and solvers. These custom-designed calibration codes are highly challenging and do not offer flexibility. Therefore, a more flexible and generic calibration tool is needed that can calibrate any code that uses Message Passing Interface/Open Multi Processing (MPI/OpenMP) for parallelization on HPC systems.

One widely used generic and independent calibration tool is the parameter estimation tool (PEST). PEST (Doherty, 2016) conducts calibration automatically based on mathematical methods and thus is applicable for optimizing nonlinear parameters. Compared with manual calibration, automatic calibration is more efficient and effective because it avoids interference from human factors (Madsen, 2000; Getirana, 2010). The uniqueness of PEST is that it operates independent of models: there is no need to develop additional programs ~~or codes~~ for a particular model except preparing the files required by PEST (as described in Sec. 3.2). PEST has four modes of operation. One of the modes is regularization mode, which supports the use of Tikhonov regularization and is found better for serving environmental models because, if implemented properly, it supports model predictions of minimum error variance, is numerically stable, and embraces rather than eschews the heterogeneity of natural systems. Singular value decomposition (SVD) can be used as a regularization device to guarantee numerical stability of the calibration problem. Parallel PEST is able to distribute many runs across many computing nodes using master-worker parallel programing. To our best knowledge, however, no approach is available that allows users to submit jobs using PEST parallelization to a typical supercomputing facility that uses job scheduling and workload management such as Simple Linux Utility for Resource Management (SLURM), Portable Batch System (PBS), and Cobalt. A previous study (Senatore et al., 2015) used PEST to calibrate WRF-Hydro over the Crati River Basin in southern Italy. Because the study area was relatively small, the authors were able to conduct the calibration using PEST in sequential mode (Alfonso ~~-~~Senatore, personal communication, 2018).

This study aims to (1) port parallel PEST to HPC clusters operated by the U.S. Department of Energy (DOE) and adapt it to work with WRF-Hydro, (2) evaluate the performance of HPC-enabled parallel PEST linked to WRF-Hydro by calibrating a flood event, and (3) explore the scale-up capability and computational benefits of HPC-enabled parallel PEST by assigning different computing resource to the entire calibration process.

2 Model description

2.1 Study area

The case presented here is one of the worst floods experienced by greater Chicago area in the past three decades; the storm occurred on April 18, 2013~~ (Campos and Wang, 2015).~~. According to the National Weather Service (NWS), the heaviest 24-hour accumulated rainfall during this storm reached 201.4, 171.1, and 136.4 mm across Illinois, Iowa, and Missouri, respectively. The Mississippi River crested at 10.8 m (1.7 m above flood stage), and the Illinois River crested in Peoria, Illinois, at 8.95 m; these river cresting broke the previous record of 8.78 m, set in 1943, and was 4.55 m above the historical normal river stage (NWS, 2013). Campos and Wang (2015) conducted three-domain nested WRF simulations to understand the dynamical and microphysical mechanisms of the event. Our study builds on the smallest domain of that study, which covers ~~the~~ Illinois, and majority of ~~Illinois,~~ Iowa~~,~~ and Missouri at a spatial resolution of 3 km (Fig. 1). The domain size is ~~750~~ ~495,000 km^2 (747 km from west to east~~ and 660~~; 657 km from south to north~~.~~).

2.2 WRF-Hydro configuration

This study employs WRF-Hydro version 5 with a basic configuration. This configuration does not use nudging techniques or spatially distributed soil-related parameters as used in the National Water Model configuration. WRF-Hydro has been tested in several different cases that focused on different hydrometeorological forecasting and simulation problems (e.g., Gochis et al., 2018; Yucel et al., 2015; Senatore et al., 2015; Arnault et al., 2016), and it shows reasonable accuracy in simulated streamflow after being carefully calibrated. For details of the WRF-Hydro modeling system, see Gochis et al. (2018). Currently, two LSMs are available in WRF-Hydro for representing land-surface column physics: Noah (Chen and Dudhia, 2001) and Noah Multiparameterization (Noah-MP; Niu et al. 2011). We utilize Noah-MP LSM because compared with Noah LSM it shows obvious improvements in reproducing surface fluxes, skin temperature over dry periods, snow water equivalent, snow depth, and runoff (Niu et al. 2011). The Noah-MP is configured at a grid spacing of 3 km, and the aggregation factor is 15; that is, starting from a 3 km LSM resolution in the domain shown in Fig. 1, hydrological routing is performed at a grid resolution of 200 m, with 3285 south-north × 3735 west-east grid cells. We use a time step of 10 seconds for the routing grid in order to maintain model stability and prevent numerical dispersion of overland flood waves. ~~The time step also meets the Courant condition criteria for diffusive wave routing on a 200 m resolution grid.~~ The WRF-Hydro is configured to be in offline or uncoupled mode—there is no online interaction between the WRF-Hydro hydrological model and the WRF atmospheric model. Overland flow, saturated subsurface flow, gridded channel routing, and a conceptual baseflow are active in this study. The gridded channel network uses an explicit, one-dimensional, variable time-stepping diffusive wave. The time step of 10 seconds also meets the Courant condition criteria for diffusive wave routing on a 200 m resolution grid. A direct output-equals-input "pass-through" relationship is adopted to estimate the baseflow. Although the baseflow module is not physically explicit, it is important because the water flow in the channel routing is contributed by both the overland flow and baseflow. If the overland flow is active as it is in this study, it passes water directly to the channel model. In this case the soil drainage is the only water resource flowing into the baseflow buckets. However, if the overland flow is deactivated but channel routing is still active, then WRF-Hydro collects excess surface infiltration water from the land model and passes this water into the baseflow bucket. This bucket then contributes the water from both overland and soil drainage to the channel flow. Therefore, the baseflow must be active if the overland flow is switched off. This study does not consider lakes and reservoirs.

We use the geographic information system (GIS) tool (Sampson and Gochis, 2018) developed by the WRF-Hydro team to delineate the stream channel network, open water (i.e., lake, reservoir, and ocean) grid cells, and groundwater/baseflow basins. Meteorological input for the WRF-Hydro model system includes hourly precipitation; near-surface air temperature, humidity, and wind speed; incoming shortwave and longwave radiation; and surface pressure. In this study, the hourly precipitation is from the National Centers for Environmental Prediction (NCEP) Stage IV analysis at a spatial resolution of 4 km. The Stage IV data is based on combined radar and gauge data (Lin and Mitchell, 2005; Prat and Nelson, 2015), and has been shown to be temporally well correlated with high-quality measurements from individual gauges (see, e.g., Sapiano and Arkin, 2009; Prat and Nelson, 2015). The other hourly meteorological inputs are from the second phase of the multi-institution North American Land Data Assimilation System project, phase 2 (NLDAS-2) (Xia et al., 2012a,b), at a spatial resolution of 12 km. NLDAS-2 is an offline data assimilation system featuring uncoupled LSMs driven by observation-based atmospheric forcing.

During the 15-day period of this studied case, light to moderate rain occurred on April 8 through 11, 2013, followed by a relatively dry period from April 12 to 15. Then a heavy rain event began on April 16 and peaked on April 18. The heaviest rain band moved east of the study area on April 19. The rainy event ended over the study area on April 20 (see Fig. S1 in Supporting Information). We start the WRF-Hydro simulation on ~~Jan.~~October 1, ~~2013~~2012, and run the model for ~~more than three~~six months to reach equilibrium. This ~~3~~6-month period is considered as spin-up time and is excluded from model calibration and evaluation. We calibrate the river discharge calculated by the WRF-Hydro model from 00UTC April 9 to 00UTC April 12, 2013, considering it long enough to achieve our objective. We then evaluate the model performance against U.S. Geological Survey (USGS) observed river discharge from 00UTC April 12 to 00UTC April 25, 2013.

3 Calibration

3.1 Platforms

We customized parallel PEST to work on three different workload managers and job schedulers: SLURM at the National Energy Research Scientific Computing Center (NERSC), PBS at the Argonne National Laboratory Computing Resource Center~~,~~ (LCRC), and Cobalt at the Argonne Leadership Computing Facility. The tests presented here are conducted on Edison and Cori at NERSC, and Bebop at Argonne LCRC, which ~~uses~~all use the SLURM workload manager and job scheduler. ~~Edison is a Cray XC30 with a peak performance of 2.57 petaflops per second, 133,824 compute cores, 357 terabytes of memory, and 7.56 petabytes of disk storage. It has 5,586 nodes and 24 cores per node.~~

The interface we have built between parallel PEST and the management software ~~(SLURM here)~~ is, in general, used for (1) setting the number of workers, and the nodes for each worker to conduct a model run (WRF-Hydro here); (2) ~~finding the nodes that are available; (3)~~ setting up the working directory for the workers; (3) finding the nodes that are available; (4) identifying the nodes that work for each worker; (5) passing the global files (same for all the working directory) to all the workers (these files include the lookup table files that are not to be calibrated, the namelist files for both LSM and hydrological sector, and restart files that generated by the previous simulations, or spin-up period); and (6) submitting the job for the entire calibration process, including parallel PEST and parallel WRF-hydro. This job can be submitted as a cold-start run or as a restart. The main difference for this interface on different management software is that different management software has its own way to ~~submit jobs and~~ identify available nodes~~. This difference requires some~~ and to submit jobs. These differences require minor changes in the ~~script~~scripts we developed, which involve finding and identifying available nodes for workers, and submitting jobs for the specific management software. See detailed comments in the published code and scripts.

3.2 PEST files and settings

PEST requires three file types in both sequential and parallel ~~mode~~modes. They are template files to define the parameters to be calibrated, an instruction file to define the format of model-generated output files, and a control file to supply PEST with the size of the problem and the settings for the calibration method. Parallel PEST uses a "master-worker" paradigm that starts model runs simultaneously by different workers (or in different folders). The master of parallel PEST communicates with each of its workers many times during a calibration. To run PEST in parallel mode, one also needs a management file to inform PEST where the working folder is for each worker and what the names and paths are for each model input file that PEST must write (i.e., lookup tables that come from template files) and each model output file that PEST must read (such as frsxt_pts_out.txt). The management file also set the maximum running time for each worker. For those workers that take longer than the maximum running time, PEST will stop the model run by that particular worker and assign that model run to another worker if there is one with nothing else to do.

To the best of our knowledge, however, parallel PEST is not designed to run on HPCs directly. We developed scripts and an interface to enable parallel PEST to run on HPCs using SLURM, PBS, or Cobalt workload managers and job schedulers. The development involved writing scripts to modify the workflow for different workload managers and job schedulers, as well as developing code to connect parallel PEST to WRF-Hydro. These developments enable parallel PEST to run many workers at the same time; each worker runs a parallel code (here WRF-Hydro) that uses more than one node, which could significantly reduce the wall-clock time of model calibrations. Although this master-worker parallelism may not be as efficient as a fully MPI approach, it is sufficient for model calibration and requires the least effort for the current parallel PEST to run on HPC systems.

This study presents calibration results from PEST using the SVD-based regularization in regularization mode to ensure numerical stability (Tonkin and Doherty, 2005). We focus on calibrating 22 parameters (see Table 1 and detail description in Sec. 3.3) using 96 observation points and 22 items of prior information for the calibrated parameters. In each item of prior information, a value equal to its default value provided by the WRF-Hydro v5.0 (or the log of its default value) is assigned for each adjustable parameter, assuming that default values are the preferred values. All prior information equations are assigned a weight of 1.0. We assigned five different regularization groups to the prior information: Manning's roughness coefficients specified by Strahler stream order in CHANPARM.TBL to one group; the parameters in HYDRO.TBL (Manning's roughness coefficients for overland flow as a function of vegetation types) to another group; and three global parameters for the Noah-MP (xslop1, refdk, and refkdt) in GENPARM.TBL to the remaining three groups. The 96 observation points are given different weights based on the inversed mean of their observed discharge during the studied period (see the detailed description in Sec. 3.3 and Sec. 4.1). For a detailed description of these settings see the PEST User Manual (Doherty, 2016).

3.3 Calibrated experiments

The primary objective of this study is to build a bridge for linking the parallel PEST and WRF-hydro on the basis of HPC clusters and to explore the computational benefits of this bridge. We do not attempt to extensively assess each individual tool or address questions in each individual domain, such as optimizing the objective functions in PEST or calibrating WRF-Hydro for a long time period considering all the relevant parameters to achieve an optimal parameter set. The calibration period thus is limited to only three days, which we believe long enough to achieve our objective and to understand WRF-Hydro's sensitivity to the calibrated parameters. We calibrated WRF-Hydro using four USGS sites (referred to as Station 1, Station 2, Station 3, and Station 4 hereafter), as shown in Fig. 1. (More USGS sites could be included if one manually reallocated the stations that were not properly assigned to the desired location on the channel network by the GIS tool.) As shown by the lower left index map in Figure 1, the study area (the red box) only covers the lower part of Upper Mississippi River Basin (UMRB) and a portion of Missouri River Basin (MORB). In order to prepare observation datasets of streamflow contributed *only* from the drainage area *within* the model domain, we identified inflows entering the model domain at three different sites, namely, sites 05411500, 06807000, and 06887500, as indicated by the black solid triangles in the index map of Figure 1. The outflows of combined UMRB and MORB can be found at the three outlets, namely, sites 07010000, 07020500, and 07022000 (named Stations 2, 3, and 4, respectively, as shown by black solid circles in Figure 1). These outlets are located sequentially at the main Mississippi River after confluence of Mississippi River and Missouri River. Thus, the observed streamflow contributed by drainage area *within* the model domain can be calculated by subtracting the sum of the discharge at the three sites (black triangles; recognized as inflow) from the discharge at each of the three outlet sites (black circles; recognized as outflow). The final derived observations of streamflow (or adjusted streamflow observation data) from the drainage area *within* this model domain are prepared for model calibration and validation. To prove this concept, we validated the consistency of the sum of observed drainage areas at inflow sites plus modeled drainage area with the overall drainage area at the outlet. The drainage area (UMRB and MORB) at outlet site 07010000 is 1.8E+12 m^2. The sum of drainage areas at three inflow sites is about 1.4E+12 m^2 (2.0E+11, 1.1E+12, and 1.4E+11 m^2 for site 05411500, 06807000, and 06887500, respectively) and the modeled drainage area is 0.36E+12 m^2; the total area is 1.76+12 m^2. This indicates that the flows from sum of three inflow sites and modeled result represent 98% of drainage area at the outflow site 07010000. Therefore, the adjusted streamflow observation data are qualified for model calibration. We then transfer the calibrated parameters to other subbasins in the study area to assess the transferability of the calibrated parameters. Although many parameters, including spatially distributed parameters and constant parameters in the lookup tables, affect the model performance, we calibrate only the parameters in lookup tables and do not consider the spatial variability of other parameters or their scaling factors. We acknowledge that some studies calibrate a single scaling factor (without considering its spatial variability, however) of overland roughness coefficients (OVROUGHRTFAC) rather than the actual value of each land type in the lookup table (e.g., Kerandi et al., 2018). Although this approach reduces the number of calibrated parameters~~, however~~, it has less flexibility because changing one factor will change all the parameters that use the same proportion.

For the calibration exercises we conduct here, the retention depth factor (RETDEPRTFAC) is fixed at 0.001. This value is reasonable because the modeled discharge of our particular configuration (Sec. 2.2) using default parameters is lower than observed discharge. Reducing this factor from 1 to 0.001 keeps less water in water ponds and more water on the surface so it can contribute to river discharge. First, we calibrate 48 parameters based on a 3-day simulation from April 9 to April 11, 2013 (Table S1 in Supporting Information). This calibration uses the estimation mode in the PEST tool and considers equal weight for all four USGS stations. We calibrate Manning's roughness coefficients for both channels and land-use types, the deep drainage (SLOPE), infiltration-scaling parameter (REFKDT), and saturated soil lateral conductivity (REFDK). Manning's roughness coefficients control the hydrograph shape and the timing of the peaks; the SLOPE, REFKDT, and REFDK control the total water volume. Second, based on the knowledge we learn from the 48-parameter calibration (see details in Sec. 4.1), for the same 3-day period, we reduce the number of calibrated parameters from 48 to 22 according to the sensitiveness of the WRF-Hydro model to the adjustable parameters. For example, during the calibration we find that Manning's roughness coefficients for several land types barely change because these land types (e.g., tundra, snow/ice) are not present in the study area. We also learn that even though the calibrated WRF-Hydro parameters can generate discharge results that closely resemble observations, the physical meaning of several parameters are not appropriate because of the wide range of those parameters that we set in the PEST control file. For example, Manning's roughness coefficient for stream order 1 (0.199) is calibrated smaller than that for stream order 2 (0.218); the overland roughness coefficients for evergreen needleleaf forest (0.043) and mixed forest (0.023) are calibrated smaller than for cropland/woodland (0.046). Neither of these is true in the real world.

We therefore adjust the range of many parameters according to the literature (Soong et al., 2012) to maintain their physical meanings (Table 1). We find that by using the same absolute weight for all four stations, the calibration helps three stations (Station 2, 3, and 4) with large water volumes to generate more reasonable results than do the default parameters; however, the results for Station 1, which has a relatively small volume of water, is not always better than the discharge that is modeled by using default parameters. Thus, we assign a weight of ~~15~~9.0 for Station 1 versus a weight of 1.0 for the other three stations according to the inversed mean of observed discharge over these four stations in April 2013. The ratio of the weights between Station 1 and the other three stations stays similar even if the means are calculated based on different time periods.

3.4 Statistics

This study employs three statistical criteria: Nash–Sutcliffe efficiency (NSE; Nash and Sutcliffe, 1970; Moriasi et al., 2007), root-mean-square error (RMSE), and Pearson correlation coefficient (PCC). RMSE and PCC evaluate model performance in terms of bias and temporal variation. NSE quantitatively describes the accuracy of modeled discharge compared with the mean of the observed data. Equation (1) calculates the NSE with defined variables:

$$NSE = 1 - \cancel{\frac{\sum_{t=0}^{n}\left(Y_t^{obs} - Y_t^{sim}\right)^2}{\sum_{t=0}^{n}\left(Y_t^{obs} - Y_{mean}^{sim}\right)^2}} \frac{\sum_{t=0}^{n}\left(Y_t^{obs} - Y_t^{sim}\right)^2}{\sum_{t=0}^{n}\left(Y_t^{obs} - Y_{mean}^{obs}\right)^2}, \qquad (1)$$

where Y_t^{obs} is the *t*th observed value from USGS sites for river discharge , $Y_t^{sim}$is the *t*th simulated value from the WRF-Hydro output, Y_{mean}^{obs} is the temporal average of USGS observed discharge, and *n* is the total number of observation time points. An efficiency of 1 (NSE = 1) corresponds to a perfect match between modeled discharge and observed data. An efficiency of 0 (NSE = 0) indicates that the model predictions are as accurate as the mean of the observed data. An efficiency below zero (NSE < 0) occurs when the model is worse than the observed mean. Essentially, the closer the NSE is to 1, the more accurate the model is.

4 Results

4.1 WRF-Hydro calibration and validation

Based on the knowledge we gained from the 48-parameter 3-day calibration, we adjust the range of critical parameters in the PEST control file to ~~main~~maintain their physical meanings. For example, we set Manning's roughness coefficient larger for stream order 1 than for stream order 2. We also adjust the parameter range of the overland roughness coefficient for multiple land covers, such as forests. We exclude the parameters that ~~WRF-Hydro~~ is not sensitive to WRF-Hydro streamflow for this study, in order to constrain the problem size considering the availability of computational resources. However, if the studied area is much larger with more land types than the study area here, then there would be more parameters to calibrate. ~~Also, hundreds~~Hundreds of constant parameters in the Noah-MP model could affect the WRF-Hydro results (Cuntz et al. 2016) and can be calibrated as well. Both these situations would increase the burden of WRF-Hydro calibration. We perform the same 3-day calibration from April 9 to April 11, 2013. Figure 2 shows the results of the 3-day modeled discharge (in cubic meters) using default and calibrated parameters after five iterations, as well as observed discharge. The four stations are calibrated by considering different weights. ~~Compared with the results~~ While the model performance for Station 1 using default and calibrated ~~by using equal weights for all~~parameters are similar, the ~~stations, by giving a higher weight to Station 1~~calibration improves the model ~~bias~~performance over ~~Station 1 is significantly reduced, with a higher NSE (0.87 with higher weight versus 0.14 with equal weight) and lower RMSE (48.1 versus 123.6). Over~~ the drainage areas represented by Stations 2, 3, and 4~~, which sit on rivers with relatively large water volumes, the~~ significantly. The modeled discharge using the default parameter underestimates the streamflow by ~~more than 65~~24-33%. PEST detects this underestimation~~ and~~, immediately adjusts the parameters and increases the modeled discharge during the first iteration. After the third iteration, the difference in calibrated results between different iterations is relatively small. We allow the PEST to conduct five iterations and use the parameters obtained from the fifth iteration as our optimum parameters. As shown in Table 2, when the optimum parameters are used, the modeled discharges are much closer to the observations ~~compared with~~than the modeled results ~~when the~~using default parameters~~ were used~~. The NSEs for the four stations increased from ~~0.73 (Station 1), -54.~~-4.8 (Station 2), ~~157.3~~-18.8 (Station 3) and -~~1316.9~~57.0 (Station 4) to 0.~~87,~~ 75, -0.~~64, 0.05~~03, and -~~58.78~~0.42, respectively, being closer to 1. It is noteworthy that, threshold values to indicate a model of sufficient quality have been suggested between 0.5 < NSE < 0.65. Here, although we see the calibration results close to the observations, the NSE are low for Stations 3 and 4. This may be because the objective function used in PEST is sum of squared weighted residuals (SSWR), which is calculated differently from NSE. Thus even if SSWR reaches a small value, the NSE might still be far from 0.5 to 0.65. Incorporating other measures into the objective function of PEST may improve the robustness of PEST calibrations. The RMSEs decreased from ~~69.3, 3925~~902.2, ~~3981~~1001.3, and ~~4391~~1399.3 m^3/sec to ~~48.1, 318.2, 308~~188.6, 228.7, and ~~934.6~~219.1 m^3/sec, respectively~~. Giving a lower weight for the three large river stations does not change the calibration results much~~.

During the validation period, compared with the modeled discharge using default parameters, as shown in Table 2, the NSEs for all four stations are increased to be closer to 1; RMSEs are significantly decreased~~ by 50% or more~~; and the correlation coefficients between the observed and modeled discharge are increased from 0.8, 0.~~76~~7, 0.~~21~~19, and 0.~~72~~65 to 0.~~98~~9, 0.~~82~~81, 0.~~80~~78, and 0.75. Compared with the results of calibration using the estimation mode (no regularization) in PEST~~,~~ (not illustrated), the SVD-based regularization generates slightly better hydrograph shape with ~~24-hour~~1-day later discharge peaks that are closer to the observations. However, a problem remains with the hydrograph shapes of the modeled discharge, especially with the modeled peak of discharge. For Station 1, the WRF-Hydro almost captures the timing of the peak of discharge, ~~although~~but it still underestimates the ~~water volume~~discharge by ~25%. ~~The reason~~One of the reasons perhaps is that this study uses a direct pass-through baseflow module, which does not account for slow discharge and long-term storage of the baseflow. Therefore, the largest contribution to river discharge is from precipitation, and groundwater does not contribute much discharge to the channels in a long-term view, as is also true for the other three large river stations. ~~Different from Station 1, for the other three large river stations, the WRF-Hydro modeled discharge increases soon after the peak of precipitation and reaches a peak on April 21, 2013, which is much earlier than the observed peak of river discharge (near April 24). The reason is that the water contributions for these stations are from a larger river basin (Mississippi River) than we included in our current study area. Thus, when a heavy precipitation event occurs over the entire river basin, there will be a significant lag time (especially at the lower part of the basin) between the peak of precipitation amount and the peak of river discharge. For example, the precipitation~~

~~over the upper part of Mississippi River Basin (MRB) has a peak amount on April 18–19, but the river discharge did not reach its peak until April 24. Because our studied area covers only half of the MRB, the modeled river discharge has a shorter delay period after the peak of precipitation than does the observed river discharge. Enlarging the study area to include the entire MRB may improve this situation. Alternatively, calibrating and validating local rivers that are included in the current study area may also reduce the bias in hydrograph shape compared to calibrating and validating large rivers. On the other hand, the WRF-Hydro simulated river discharge decreases soon after it reaches the peak and much earlier than the observed discharge. The reason is again that the direct pass-through baseflow employed by this study does not account for slow discharge and long-term storage of the baseflow.~~ As a result, the contribution from the baseflow to the river discharge in model simulations does not stay as long as in real situations. In the observations, the river discharge decreases from the peak at a speed of ~500 m^3/sec per day, while the modeled river discharge decreases from the peak at a speed of ~1667 m^3/sec per day. Using exponential storage-discharge function for the baseflow may improve this situation. Other reasons include that the parameter range we set in the PEST control file is perhaps not wide enough, as we can see from Table 1 that, several optimal parameters hit the bound of parameter ranges. Allowing wider parameter ranges may improve the calibration results.

Alternatively, instead of calibrating the stations that have large drainage area and water coming from outside of the current model domain, we have also tested calibrating small flows at local stations that have relatively small drainage area covered by the current study area. This requires to generate a new high-resolution GIS data file to distribute the stations of interest. We first run the WRF-Hydro model for 6 month using default parameters to spin up the model, and then we calibrate the model based on observations of these local stations. Results including figures and tables are shown in Supporting Information. The calibration results are improved compared to the results that use default parameters, although further improvements are still needed. This again may be because the parameter range are not wide enough to consider the possible values of parameters that work for these specific areas represented at local stations, as we see many optimal parameters hit the bound of the parameter range. More tests to figure out a better set of parameters are needed for future investigation, which is beyond the scope of this study, since our goal is to present the feasibility of HPC enabled PEST.

4.2 Computational benefits of parallel PEST on HPCs

The ability to scale up the calibration of WRF-Hydro by using parallel PEST on HPC systems is determined by two factors: the scale-up capability of parallel PEST and the scale-up capability of WRF-Hydro. In calibrating WRF-Hydro, PEST first makes as many model runs as there are adjustable parameters to calculate Jacobian matrix (Doherty, 2016). The Jacobian matrix has a column for each calibrated parameter and a row for each observation and each item of prior information that set in the PEST control file. These model runs are independent between workers and can be easily parallelized. Each worker runs the model with temporarily incremented parameters that are defined in the template and control files. Then, PEST needs to make additional model runs to test parameter updates. Different from the Jacobian runs, these additional runs are performed by using different Marquardt lambdas, and the search for a Marquardt lambda that achieves the best set of parameters is a serial iterative process. The lambda to use for the next run depends on the outcome of the model run conducted using the previously chosen lambda. Although serial testing of Marquardt lambdas may quickly find the optimal Marquardt lambda in the first or second series of model runs, it is an inefficient use of computing resources because other processors are idle while only one process is searching the lambdas. This is especially true when the model domain is large and requires extensive computing resources. This study employs "partial parallelization" for the lambda-testing procedure (Doherty, 2016), so multiple workers can be used to calculate parameter upgrades based on a series of lambda values that are related to each other by a factor of RLAMFAC set in the PEST control file. We also set the value of PARLAM to -9999 in the management file so only one cycle of parallel WRF-hydro runs is devoted to testing Marquardt lambdas. For additional details on these parameters and their settings see the PEST User Manual (Doherty, 2016).

In this study we test the computational performance of HPC-enabled parallel PEST using different number of workers (6, 12, and 23) for the 22-parameter calibration. As shown in Table 3, we conducted ~~five~~six experiments: Test 1 uses 23 workers, Test 2 uses 12 workers, and Test 3 uses 6 workers. All three tests use two nodes for each worker to run WRF-Hydro in parallel. The maximum number of lambda-testing runs undertaken per iteration is set to 15, 10, and 5 for

~~Test~~Tests 1, 2, and 3, respectively, to ~~make sure~~assure that only one cycle of WRF-hydro runs is devoted (using 15, 10 and 5 workers from Tests 1, 2, and 3, respectively) to testing Marquardt lambdas. Note that the maximum number of lambda-testing runs should be set equal to or less than the workers available. Otherwise, another cycle of WRF-hydro runs needs to be conducted. In fact, generating more Marquardt lambdas does not always guarantee that the best Marquardt lambdas are generated. In contrast, it may make the model convergence slower (here, PEST) or even model failure.

In order to test the trade-offs between the computing nodes used for running parallel WRF-Hydro and the workers used for running parallel PEST, Tests 4~~ and 5~~, 5 and 6 use the same number of workers (six) as Test 3 but use different number of nodes for each worker to run WRF-Hydro in parallel. Explicitly, Test 4 uses four nodes per worker, ~~and~~ Test 5 uses six nodes per worker~~. Both tests use six workers for running the parallel PEST.~~, and Test 6 uses eight nodes per worker. The maximum number of lambda-testing runs undertaken per iteration is set to five for ~~both~~ Tests 4, 5 and ~~5~~6. Note that the time costs in Table 3 are limited to only one iteration. Conducting more iterations will increase the cost of wall-clock time and computing resource, but will not change the conclusion for the scale-up capability and computational benefits for HPC-enabled parallel PEST linked to WRF-hydro.

PEST needs to run the WRF-Hydro model at least as many times as the number of calibrated parameters (22 here). In fact, PEST runs the model 23 times in the first round (or the first iteration) with initial parameter values and for the first Jacobian matrix. From the second iteration, it runs the model 22 times to calculate Jacobian matrix. Therefore, if there are fewer than 23 workers, the time cost for the first round of Jacobian matrix calculation will increase accordingly. For example, as shown in Fig. 4a, when we assign 12 (and 6) workers to parallel PEST, the time cost for calculating the Jacobian matrix is increased by a factor of 2 (and 4) compared with the time cost of using 23 workers. The time cost for the parameter upgrade stays similar for the three experiments because only one cycle of WRF-hydro simulation is conducted to test the Marquardt lambdas. As a result, the total time cost for Test 2 is ~1.5 times more than that for Test 1, and the total time cost for Test 3 is ~1.5 times more than that for Test 2 (Fig. 4b). By extrapolating the speedup curve shown in Fig. 4a and Fig. 4b, we expect the total time cost to be ~1516 minutes when using only one worker (or sequential mode), which is about 15 times slower compared with running the PEST in parallel mode using 23 workers. For this particular study with 22 adjustable parameters, we expect the time cost most likely to stay the same even if one increases the number of workers to more than 23, because PEST runs WRF-Hydro only 23 or 22 times for each iteration. Assigning more workers for this particular study would most likely render some workers idle and is not an efficient use of computing resources. PEST may run WRF-Hydro more than 22 times (e.g., 44 times) if higher-order finite differences are employed. In this case, assigning more workers (e.g. 45 workers) may further speed up the calibration process. On the other hand, for the same case study and using the same number of nodes for running parallel WRF-Hydro, we can estimate the computing speedup by assuming an increase in the number of calibrated parameters to 50. This would be the case, for example, to evaluate model sensitiveness to the physics in Noah-MP or the spatial variabilities of certain parameters. We then expect to use 51 workers to ~~achieve~~calculate the ~~best computing performance for parallel PEST~~Jacobian matrix in only one cycle. This would then be 28–30 times faster than running PEST using one worker (or in sequential mode). Similarly, if 100 parameters were used for the calibration for the same case study, a factor of up to 60 speedup in the calibration process would be achieved by running HPC-enabled parallel PEST.

In addition, by increasing the number of nodes for each worker to conduct WRF-Hydro (Tests 3, 4, 5, and ~~5~~6), the time cost for the entire calibration process is significantly reduced (Figs. 4c and 4d). Specifically, the WRF-hydro scales up well when using four, six, and ~~six~~eight nodes compared with using two nodes per worker for running the WRF-Hydro. Both the time spent on calculating the Jacobian matrix and the time spent on testing the parameter upgrades are decreased by 49% ~~and ~~67%, and 77%, respectively, when using four, six, and ~~six~~eight nodes. Therefore, the total time spent is also decreased when using more nodes for each worker (see Table 3). ~~Increasing the number of nodes to eight for each worker will most likely further decrease the time cost by 70–75% compared with using only two nodes per worker.~~ Moreover, if one has a larger study area such as the entire contiguous United States, we expect the WRF-Hydro to have an even better scale-up capability (e.g., on dozens of nodes) than this study. ~~Overall, based on the experiments we conduct here, using 23 workers for parallel PEST and six nodes for each worker to run parallel~~

~~WRF-Hydro would cost the least wall-clock time—about 32 min for one iteration for this particular study.~~

While these numbers in Table 3 and Figure 4 are helpful to demonstrate the scale-up capability of each component (PEST and WRF-Hydro), they do not answer questions such as, if one has certain number of nodes, how many workers and how many nodes per worker should be used to achieve the highest efficiency of the WRF-Hydro calibration using HPC-enabled PEST? On the other hand, one may have unlimited computational resource, but would like to complete the calibration in a short time period. We present scalability analysis below to answer these questions. First, we generate more scenarios using different number of workers and nodes per worker by extrapolating the existing time and computing costs based on the experiments that are already conducted. These scenarios use 23 or 12 workers, and 4, 6, or 8 nodes per worker, respectively. Since we have conducted simulations using the same number of nodes per worker, the cost for these scenarios are easily predicted.

As shown in Figure 5, compare with Test 3 (which requires the least computing resource —12 nodes in total), having more workers (with the same number of nodes for each worker, e.g., Tests 1 and 2), takes more time than the ideal curve. The ideal curve assumes a linear speedup based on the time cost of Test 3. However, using the same number of workers and increasing the number of nodes for each worker (e.g., Tests 4, 5, and 6) can achieve the ideal speedup. Even when using 12 workers, increasing the number of nodes for each worker can still achieve a speedup close to the ideal curve. Using 23 workers will not achieve the ideal speedup. Therefore, if one only has a certain number of nodes available, we recommend to use relatively small number of workers but large number of nodes for each worker. For example, if one has 48 nodes, then there are three options can be considered: using 23 workers and 2 nodes per worker; 12 workers and 4 nodes per worker, and 6 workers and 8 nodes per worker. Other partition (16x3; or 8x6) between numbers of workers and nodes per worker are not as efficient as above. These three options will cost 103, 72 and 60 min, respectively, to finish one iteration. Thus, using 6 workers and 8 nodes per worker is the most efficient way to consume the limited computing resource. On the other hand, if one would like to conduct the calibration in a short time period without any limits for the computing resource, then using 23 workers and 8 nodes (perhaps even more nodes depending on the size of the model domain and the scale up capability of WRF-Hydro), will finish one iteration in ~24 min.

4.3 Evaluation of spatial transferability of the calibrated parameters

To assess the transferability of the calibrated parameters, we apply the optimum parameters obtained from the calibration for the four stations (black circles) in Fig. 1 to another set of four stations (crosses in Fig. 1) in the study area. All four sites are located on relatively small rivers, so the lag time between precipitation peak and the discharge peak are much shorter than that for the stations on the lower part of MRB (e.g., Stations 2, 3, and 4). The assessment compares the observed discharge with the closest grid cells from the discharge output of WRF-Hydro. Figure ~~5~~6 shows the observed and modeled discharge using default and the optimum parameters. Overall, WRF-Hydro’s default parameters underestimate the discharge and misrepresent the timing of discharge peaks compared with observations over the four assessed stations (Stations 5, 6, 7, and 8). By using the calibrated parameters from other sites over the area, the model results increase the discharge and shift the hydrograph shape so they are much closer to the observations than model results using default parameters. The absolute error of simulated discharge decreases by 13.1%, 38.3%, and 71.6%, respectively, over Stations 6 through 8 (Station 5 shows a 6% increase of absolute error), compared with the default simulated discharge. We also find that using the SVD-based regularization for the PEST calibration captures the timing of discharge peak better than using the estimation mode, which is one-day earlier than the observations reaching the discharge peak.

5 Summary and discussion

WRF-Hydro is a new, and perhaps the first practical, computer code that can run on HPC systems and can model the entire hydrological cycle using physics-based submodels and high-resolution input datasets (e.g., radar). The hydrological community has desired this capability for decades, although it requires intensive computing resources. Thus, the calibration of this model would ideally be conducted on HPCs in parallel as well, especially when the model covers a large domain rather than the basin scale. This study ports an independent model calibration tool, parallel PEST, to HPC clusters and links it to WRF-Hydro to help WRF-Hydro users calibrate the model within a much shorter wall-clock time period. The bridge we build here (between parallel PEST and WRF-Hydro on the basis of HPC systems) can be applied to any other hydrological models and Earth system models that use parameterizations to represent model physics. We present the operational feasibility of the HPC-enabled parallel PEST by evaluating the performance of calibrated WRF-Hydro against observation in hydrograph features such as volume and timing of flood events. We examine the scale-up capability and computational benefits of the tool by assigning different computing resource for PEST and for WRF-Hydro. While this study presents the optimum parameters identified from the calibration of the particular flood event, the parameters can be significantly different if one uses different physics, such as exponential storage-discharge function for a groundwater model or reach-based channel routing. Our preliminary testing shows that using exponential storage-discharge function with the default parameters provided by WRF-Hydro, the modeled discharge was larger than that of observations. Thus, the calibration will need to adjust the parameters to reduce the discharge. Our study finds that for calibrating 22 parameters, using the same computing resource for running WRF-hydro, the HPC-enabled PEST calibration tool can speed up WRF-Hydro calibration by a factor of 15, compared with running PEST in sequential mode. The speedup factor can be larger when the number of parameters needing calibration is higher (e.g., 50 or 100).

The following are several key points that we would like to mention to inform future studies:

1. In this study, we consider using the prior or regularization information only for the parameters that we calibrate. As is the case with solving inverse problems, prior information is added to improve the smoothness of the solutions. In order to build a more comprehensive calibration, an important aspect that can be considered is to enrich the prior with the available historical data. For example, in this particular case, one can use the historical observation data (e.g., April and May from the past few years) to enrich the prior information for the parameters. Hence, the regularization objective function in PEST will constitute not only the discrepancies between parameters and their “current estimates” but also the discrepancies between WRF-Hydro simulations and preferred values (which is the observed time series of historical discharge). Additionally, one can use the pilot points technique described by Doherty (2005) in conjunction with parameter estimation to add

more flexibility to the calibration process. This will be potentially beneficial in improving the predictions.

2. To focus on our main goal, we calibrate only the parameters in lookup tables. However, we acknowledge that using a single value to represent a physics for a large domain could be problematic, especially we expect the HPC-enabled parallel PEST to execute with WRF-Hydro for large domains. This situation often needs parameter regionalization. For example, WRF-Hydro ~~v5~~version 5.0 has many spatially distributed parameters available, such as OVROUGHRTFAC— the overland flow roughness scaling factor ~~(OVROUGHRTFAC),~~, RETDEPRTFAC— the factor of maximum retention depth ~~(RETDEPRTFAC),~~, and the soil-related parameters (when compiled with SPATIAL_SOIL=1). Calibrating these spatial parameters based on grid scale (e.g., catchments) rather than a single value will give the model more flexibility and thus better fit the observations (Hundecha and Bardossy, 2004; Wagener and Wheater, 2006). In practice, for example, one can include regional OVROUGHRTFACs (e.g., their lower/upper bounds, and default values) in the PEST control file based on catchments. However, the selection of the locations and sizes of catchment may introduce significant uncertainties to the calibration results, which require systematic and comprehensive investigation and understanding of the study area.
3. This study is limited to calibrating the observed streamflow only based on the format of one of WRF-Hydro model outputs for individual station or point (frxst_pts_out.txt). It is feasible, however, to calibrate other variables as long as the observation data is available. For example, one can either find the closest point from the gridded dataset to the observation location and then compare that model grid to observations; or one can change the WRF-Hydro input/output code to output other variables in the frxst_pts_out.txt file, so they can still use the same interface we developed here to calibrate other variables ~~instead~~ in addition to the discharge.
4. The optimal parameter set obtained from this study is from the 5th iteration of parallel PEST by testing five Marquardt lambdas. Testing different number of lambdas or calibrating different number of parameters may generate a different set of optimal parameters. These parameter sets can all make physical sense and be equally good for reproducing observed discharges. This problem is named equifinality (Beven and Freer,

2001; Savenije, 2001), which is an important source of model uncertainty. To reduce the model uncertainty through reducing the equifinality, hydrologists carry out additional modelling objective for model evaluation to find more useful parameter sets (Mo and Beven, 2004; Gallart et al., 2007). Alternatively, inspired by No. 3 discussed above, one can calibrate the WRF-hydro model based on more than one variables, such as discharge and soil moisture (or heat flux or water table depth) to reduce the number of optimal parameter sets, and thus reduce the model uncertainty of predictions for these variables.

5. While this study ported the parallel PEST to HPC system and linked it to WRF-Hydro, we note that BEOPEST is available in the PEST family. BEOPEST has the same functionality as parallel PEST but uses a different approach for communication between master and workers. Working with HPC-enabled BEOPEST may save total time cost since BEOPEST uses the Transmission Control Protocol ~~(TCP) ~~and the Internet Protocol~~ (IP)~~ instead of message files (reading input and writing output between master and works) for communication. We expect it to be relatively straightforward to use BEOPEST to calibrate WRF-hydro on HPCs since the interface remains ~~similar~~the same, except one needs to copy the template and instruction files in addition to the global files (see Section 3.1) into each working folder.

Data and Code availability. The observed river discharge is downloaded from the USGS Surface-Water Data website, available at https://waterdata.usgs.gov/nwis/sw. The Stage IV precipitation data were downloaded from https://data.eol.ucar.edu/dataset/21.093. PEST was downloaded from http://www.pesthomepage.org/Downloads.php. We use the Unix PEST version 13.6. The scripts and files that are developed in this study and required by PEST for calibrating WRF-Hydro are available at http://doi.org/10.5281/zenodo.3247116~~http://doi.org/10.5281/zenodo.2588506~~.

Author contributions. JW proposed the project and developed the study case in WRF and WRF-Hydro. CW developed the scripts/code to port the parallel PEST to DOE supercomputers and adapt it to work with WRF-Hydro. VR provided important input for the regularization calibration method. AO operated the ArcGIS tool to delineate the high-resolution grid cells to include stream channel network, open water, and groundwater/baseflow basins. ~~RK provide~~EY provided important input for hydrology during the revision of this manuscript. RK provided high-level guidance and insight for the entire project. All authors commented on this manuscript.

Competing interests. The authors declare that they have no conflict of interest

Acknowledgments. This work is supported under a Laboratory Directed Research and Development (LDRD) Program at Argonne National Laboratory, through U.S. Department of Energy (DOE) contract DE-AC02-06CH11357. Computational resources are provided by the DOE-supported National Energy Research Scientific Computing Center, Argonne National Laboratory Computing Resource Center, and Argonne Leadership Computing Facility. Our special thanks to the PEST developers and entire WRF-Hydro team, especially Kevin Sampson for his guidance on the ArcGIS tool. We gratefully thank the two reviewers for their valuable comments and suggestions, which tremendously improved this manuscript.

References

Arnault, J., Wagner, S., Rummler, T., Fersch, B., Bliefernicht, J., Andresen, S., and Kunstmann, H.: Role of runoff–infiltration partitioning and resolved overland flow on land–atmosphere feedbacks: A case study with the WRF-Hydro coupled modeling system for West Africa, J. Hydrometeorol., 17, 1489–1516, 2016.

Beven, K., and Freer, J.: Equifinality, data assimilation, and uncertainty estimation in mechanistic modelling of complex environmental systems using the GLUE methodology, J. Hydrol., 249, 11-29, 2001.

Campos, E., and Wang, J.: Numerical simulation and analysis of the April 2013 Chicago Floods, J. Hydrol., 531, 454–474, 2015.

Chen, F. and Dudhia, J.: Coupling an advanced land surface-hydrology model with the Penn State-NCAR MM5 modeling system, Part I: Model implementation and sensitivity, Mon. Weather Rev., 129, 569–585, 2001.

Cuntz, M., Mai, J., Samaniego, L., Clark, M., Wulfmeyer, V., Branch, O., Attinger, S., and Thober, S.: The impact of standard and hard-coded parameters on the hydrologic fluxes in the Noah-MP land surface model, J. Geophys. Res. Atmos., 121, 10,676–10,700, doi:10.1002/2016JD025097, 2016.

Doherty, J.: PEST: Model Independent Parameter Estimation, User Manual, 6th ed., Watermark Numerical Computing, Brisbane, Queensland, Australia, 2016.

Doherty, J.: Ground water model calibration using pilot points and regularization, Groundwater, 41(2), 170–177, 2005.

Gallart, F., Latron, J., Llorens, P., and Beven, K. J.: Using internal catchment information to reduce the uncertainty of discharge and baseflow predictions. Adv. Water Resour. 30(4), 808–823, 2007.

Getirana, A. C. V.: Integrating spatial altimetry data into the automatic calibration of hydrological models, J. Hydrol., 387 (3-4), 244–255, doi: 10.1016/j.jhydrol.2010.04.013, 2010.

Gochis, D. J., Barlage, M., Dugger, A., FitzGerald, K., Karsten, L., McAllister, M., McCreight, J., Mills, J., RafieeiNasab, A., Read, L., Sampson, K., Yates, D., and Yu, W.: The WRF-Hydro modeling system technical description, (Version 5.0). NCAR Technical Note. 107 pages. Available online at:
https://ral.ucar.edu/sites/default/files/public/WRFHydroV5TechnicalDescription.pdf, 2018.

Hundecha, Y., and Bárdossy, A.: Modeling of the effect of land use changes on the runoff generation of a river basin through parameter regionalization of a watershed model, J. Hydrol., 292, 281–295, 2004.

Kerandi, N., Arnault, J., Laux, P., Wagner, S., Kitheka, J., and Kunstmann, H.: Joint atmospheric-terrestrial water balances for East Africa: A WRF-Hydro case study for the upper Tana River basin, Theor. Appl. Climatol., 131, 1337–1355, doi: 10.1007/s00704-017-2050-8, 2018.

Lin, Y., and Mitchell, K. E.: The NCEP stage II/IV hourly precipitation analyses: Development and applications, Preprints, 19th Conf. on Hydrology, San Diego, CA, Amer. Meteor. Soc., 1.2., 2005.

Madsen, H.: Automatic calibration of a conceptual rainfall–runoff model using multiple objectives, J. Hydrol., 235, 276–288, 2000.

Mo, X., and Beven, K.: Multi-objective parameter conditioning of a three-source wheat canopy model. Agricultural & Forest Meteorol. 122(1–2), 39–63, 2004.

Moriasi, D. N., Arnold, J. G., Van Liew, M. W., Bingner, R. L., Harmel, R. D., and Veith, T. L.: Model evaluation guidelines for systematic quantification of accuracy in watershed simulations, Transactions of the ASABE, 50 (3), 885–900, 2007.

Nash, J. E., and Sutcliffe, J. V.: River flow forecasting through conceptual models, part I – A discussion of principles, J. Hydrol., 10(3), 282–290, doi: 10.1016/0022-1694(70)90255-6, 1970.

Niu, G.-Y., Yang, Z.-L., Mitchell, K. E., Chen, F., Ek, M. B., Barlage, M., Kumar, A., Manning, K., Niyogi, D., Rosero, E., Tewari, M., and Xia, Y.: The community Noah land surface model with multiparameterization options (Noah-MP): 1. Model description and evaluation with local-scale measurements, J. Geophys. Res., 116, D12109, doi: 10.1029/2010JD015139, 2011.

NWS (National Weather Service): Record river flooding of April 2013, https://www.weather.gov/ilx/apr2013flooding, 2013.

Prat, O. P., and Nelson, B. R.: Evaluation of precipitation estimates over CONUS derived from satellite, radar, and rain gauge data sets at daily to annual scales (2002-2012), Hydrol. Earth Syst. Sci., 19, 2037–2056, doi: 10.5194/hess-19-2037-2015, 2015.

Sampson, K., and Gochis, D.: WRF Hydro GIS Pre-processing tools, Version 5.0 Documentation, 2018.

Sapiano, M. R. P., and Arkin, P.A.: An intercomparison and validation of high-resolution satellite precipitation estimates with 3-hourly gauge data, J. Hydrometeor., 10, 149–166, doi: 10.1175/2008JHM1052.1, 2009.

Senatore, A., Mendicino, G., Gochis, D. J., Yu, W., Yates, D. N., and Kunstmann, H.: Fully coupled atmosphere-hydrology simulations for the central Mediterranean: Impact of enhanced hydrological parameterization for short and long time scales, J. Adv. Model. Earth Syst., 7(4), 1693–1715, doi: 10.1002/2015MS000510, 2015.

Savenije, H. H. G.: Equifinality, a blessing in disguise?, Hydrol. Process., 15, 2835-2838, 2001.

Soong, D. T., Prater, C. D., Halfar, T. M., and Wobig, L. A.: Manning's roughness coefficients for Illinois streams, U.S. Geological Survey Data Series 668, 2012.

Tonkin, M. J., and Doherty, J.: A hybrid regularized inversion methodology for highly parameterized environmental models, Water Resource Research, 41, W10412, doi:10.1029/2005WR003995, 2005.

~~Wagenera~~Wagener, T., and Wheater, H. S.: Parameter estimation and regionalization for continuous rainfall-runoff models including uncertainty, J. Hydrol., 320, 132–154, 2006.

Xia, Y., Mitchell, K., Ek, M., Sheffield, J., Cosgrove, B., Wood, E., Luo, L., Alonge, C., Wei, H., Meng, J., Livneh, B., Lettenmaier, D., Koren, V., Duan, Q., Mo, K., Fan, Y., and Mocko, D.: Continental-scale water and energy flux analysis and validation for the North American Land Data Assimilation System project phase 2 (NLDAS-2), 1: Intercomparison and application of model products, J. Geophys. Res., 117, D03109, doi: 10.1029/2011JD016048, 2012a.

Xia, Y., Mitchell, K., Ek, M., Cosgrove, B., Sheffield, J., Luo, L., Alonge, C., Wei, H., Meng, J., Livneh, B., Duan, Q., and Lohmann, D.: Continental-scale water and energy flux analysis and validation for the North American Land Data Assimilation System project phase 2 (NLDAS-2). 2. Validation of model-simulated streamflow, J. Geophys. Res., 117, D03110, doi: 10.1029/2011JD016051, 2012b.

Yucel, I., Onen, A. Yilmaz, K. K., and Gochis, D. J.: Calibration and evaluation of a flood forecasting system: Utility of numerical weather prediction model, data assimilation and satellite-based rainfall, J. Hydrol., 523, 49–66, 2015.