Workflow for DREAMzs Calibration of RHESSys – In Progress

0. **Download data from the SA**
   Filename:

1. **Evaluate parameter correlations and select parameters to calibrate.**
   The RHESSysParamSelection.R script does the following. This script is run using the RData file that is output after running the step 36 file.

   Evaluate Parameter Correlations
   Cluster analysis using the Chu and Hahn (2009) metric is used. Dendrograms with different cutoff levels are evaluated. Based on these results, parameter groupings were determined to be infeasible because parameters form groups with different parameters in them, and so clustering would not help reduce dimensionality.

   Select the Parameters to Calibrate
   The top 10% of the EEs for each of the 6 metrics are gathered for each of the basin and hillslope evaluations. All parameters whose 95th percentile EE estimates were greater than the 10% threshold value were also selected. This resulted in a total of 42 parameters. Other users can specify a different % threshold to use.

2. **Define priors for the parameters to be calibrated**
   Each parameter to be calibrated is assumed to have a uniform prior distribution, with bounds based on literature values or physical laws. The same priors were used as the SA, except for impervious landuse (set spatially from land cover dataset for l4 and l3 parameters), septic water load lower bound was decreased to allow for 0 septic load, and Ksat parameters' lower bounds were decreased slightly to allow for constraints to be met.

   The parameter bound file created in step 6 was updated with these new bounds, and a new file was created. This is BaismanCalibrationParameterProblemFile.csv in this study. Only the RHESSys parameters are placed in this file, not the likelihood parameters.

3. **Select MCMC Chain Starting Locations**
   The R package lhs is needed for this. The AnalyzeLikelihoods.R script was used to select the chain starting locations using a Latin Hypercube Sample (improvedLHS function in R) from the RHESSys parameter hyperspace defined by the bounds in the previous step. 4 sets of 10 chains, each with a different random seed, were used to evaluate random seed variability in the MCMC search. Files with the chain starting locations are named BaismanChainStarts_LHS10.txt.

   All parameters that were not calibrated were fixed at the same values for all chains. Most fixed values were the same as were used as defaults in the sensitivity analysis. The new values are in

Commented [SJD(1]: R package

modified GIS2RHESSys files vegCollection_modified_Cal.csv, lulcCollectionEC_Cal.csv, and soilCollection_Cal.csv (These files were made by editing the step 14 files).

4. **Check and adjust parameter values to meet constraints**
   Parameter values were checked for satisfying the constraints, and adjusted if necessary using the DREAM_ParameterBoundChecks_ChainStarts.py script. This script is a modified MorrisSampling.py script to account for only the parameters that are being calibrated. See step 7 for possible edits to make for this script. The RunChainStarts.sh script is used to run this file (see directory structure in step 41 for where to run this file). The following commands are needed:

   - working directory (e.g., RHESSysRuns)
   - initial random seed for the chain
   - directory of def files for calibration
   - round tolerance (<= 10)
   - problem file name with extension (e.g., BaismanCalibrationParameterProblemFile.csv)
   - chain parameter sample text file name without extension (e.g., 'BaismanChainStarts', 'BaismanChain_1' where 1 is chain iteration)

5. **Check that R packages for MCMC calibration are installed**
   The following packages are needed: BayesianTools, parallel, foreach, iterators, doParallel, and rlist

Commented [SJD(2]: R packages. BayesianTools was modified greatly.

6. **Setup files on Rivanna for calibration**
   The following directories (and files) should be made (placed) inside of a main folder (e.g., Baisman30mDREAMzs-10Ch):

   RHESSysRuns folder:

   output folder – empty

   RunChainStartsCheck.sh – run in step 39

   DREAM_ParameterBoundChecks.py, DREAM_ParameterBoundChecks_ChainStarts.py

   BaismanChainStarts_LHS10.txt – from step 38

   BaismanCalibrationParameterProblemFile.csv – from step 37

   MakeDefs_fn_Chains.py

   ModifyVeg.py

   RHESSysDREAM_NoG2W_rr_arg

   RHESSys_Baisman30m_g74 folder:

- defs
    - all def files, with constants at their assumed values. From step 38.
- tecfiles
    - tec_daily_cal.txt
- output - empty
- clim
    - Cal_Feb2020Revised.tmax
    - Cal_Feb2020Revised.tmin
    - Cal_Feb2020Revised.rain
    - Cal_Feb2020Revised.base
- flows – same as output from GIS2RHESSys in step 4
    - subflow.txt
    - surfflow.txt
- worldfiles – same as output from GIS2RHESSys in step 4
    - worldfile.csv
    - worldfile.hdr – change the climate prefix name if needed

TNFun folder:

- WRTDS interpolation tables
- WRTDS_modifiedFunctions.R from step 19

GIS2RHESSys folder – same as step 2

LikelihoodFun folder:

- likelihood.py from step 25
- TN_MLEfits_Cal.py
- Flow_MLEfits_Cal.py

obs folder:

- The code will trim these two files to the specified start and end dates:
- TN_Feb2020Revised_Cal.txt – from step 31
- BaismanStreamflow_Feb2020Revised_Cal.txt – from step 31

RunRHESSysDREAMCal-10Ch_rr_arg.sh

RunJoinDREAM.sh

JoinDREAM.R

Finally, one directory higher than the main folder, place the MoveDREAMFiles_10Ch.sh script.

7. Run DREAM MCMC Calibration

Commented [SJD(3]: Include note about restarts of the algorithm.

The RunRHESSysDREAMCal-10Ch_rr_arg.sh script is used to run the RHESSysDREAM_NoG2W_rr_arg.R script that sets up and runs the DREAM MCMC calibration. The following edits may be needed to run the shell script:

- SLURM commands (directories). Number of cores should be number of chains + 1

The rest of the commands are explained in the shell script. Some additional details and relations to steps in this workflow are provided here:

- #2 – R random seed should be different when DREAM is restart.
- #3 - WRTDS script from step 19
- #4 - Observed streamflow record from step 31
- #5 - Observed TN record path from step 31
- #10 -16 - WRTDS interpolation tables
- #17 – Full path to the file that describes the parameter names and bounds from step 27.
- #18 - File with chain starting locations (.csv file from step 39) OR the output from a previous chain run (.RData file)
- #20 – iterations = (Desired iterations per chain)*(number of chains)
- #21 - #30 – additional DREAMzs R function parameters.
- #34 - File that describes the parameter names and bounds from step 27. Just the name, not the full path.
- #40 – 42 - Initial random seed values. These same values are used even upon restart of DREAM.
- #43 – 44 - paths to the processed streamflow and TN observation .txt files. Processing happens within this R script, and these files are written in this script.
- #45 – 46 - Number of initial locations for the multi-start MLE solver from step 26

Note: 100 chain steps with 4 sets of 10 chains = about 111,000 files and 60 GB space (not RAM) in 3.5 days of wall clock time

8. Join the output files into fewer files

The RunJoinDREAM.sh script is used to run the JoinDREAM.R script to join files and summarize useful output information. The following edits to the shell script may be needed:

- SLURM commands (e.g., directories, email)
- System argument #6: The starting value for the chain step index will change for each successive DREAM run.

9. **Move files to permanent storage**
   Run the MoveDREAMFiles_10Ch.sh script. The following edits may be needed:
   - Directories (SLURM, permanent storage directory, directory where files to be moved are located)
   - The chain step indicator: (e.g., s100). Do a find and replace for _s###. The ### should be the last step in the chain from the completed DREAM run.

10. **Analyze MCMC Output**
    The PlotRHESSysDREAM.R script is used to visualize the MCMC output results and to compute summary convergence metrics.

11.

Script File Descriptions in Use Order

1. RHESSysParamSelection.R: Used to select parameters for calibration based on their EEs.
2. AnalyzeLikelihoods.R: Script to evaluate the likelihood of parameter sets and to select the MCMC chain starting locations based on those likelihoods.
3. DREAM_ParameterBoundChecks_ChainStarts.py: Script to check and adjust (if needed) the parameter constraints for the MCMC chain starting locations.
4. RunChainStartsCheck.sh: Script used to run the previous file.
5. DREAM_ParameterBoundChecks.py: Script to check and adjust (if needed) the parameter constraints for the MCMC chain proposal locations.
6. MakeDefs_fn_Chains.py: Script to make RHESSys def files for the MCMC chain runs with the calibration parameters.
7. Flow_MLEFits_Cal.py, TN_MLEFits_Cal.py: Scripts used during calibration for maximum likelihood estimation of the likelihood function parameters for flow and TN.
8. RHESSysDREAM_NoG2W_rr_arg.R: Script to setup and run the DREAM MCMC calibration.
9. RunRHESSysDREAMCal-10Ch_rr_arg.sh: Shell script to run the previous file.
10. JoinDREAM.R: Script to join the output from the completed DREAM run into fewer files.
11. RunJoinDREAM.sh: Shell script to run the previous file.
12. MoveDREAMFiles_10Ch.sh: Script to move DREAM output files to permanent storage directories.
13. PlotRHESSysDREAM.R: Script to plot and analyze DREAM output.