

User's Manual for The Hydrological Emulator (HE, version v1.0.0)**Yaling Liu**cauliuyaling@gmail.com

January 2018

NOTE: Please contact the developer Dr. Yaling Liu (cauliuyaling@gmail.com) before use and distribution of the Hydrological Emulator (HE), this is to help us track its use among the community and improve its capability according to the user's needs.

Questions and reports on bugs are welcome.

Table of Contents

1. Overview of the HE	2
2. Structure of the HE	3
3. Emulation by using the HE.....	4
4. Main source codes.....	5
4.1 Codes tree of main programs	5
4.1.1 Calibration process	5
4.1.2 Validation process	6
4.2 Codes and data availability	7
5. Input data	7
5.1 Climate data	7
5.2 Benchmark runoff product	8
5.3 Baseflow Index (BFI)	9
5.4 Input data organization	9
6. Output data	11
6.1 Differences in the outputs from calibration and validation	11
6.2 Output data organization	11
7. How to run the HE.....	14
Reference	15

1. Overview of the HE

The hydrological emulator (HE) is an open-source and easy-to-use tool that can efficiently mimic complex global hydrological models (GHMs). The HE is featured by: 1) minimum number (only 5) of parameters; 2) minimal climate input that is easy to acquire; 3) simple model structure; 4) reasonable model fidelity that captures both spatial and temporal variability; 5) extraordinary computational efficiency; 6) applicable in a range of spatial scales; and 7) open-source and well-documented. It can be used to assess variations in water budgets at a variety of spatial scales of interest (e.g., basin, region or globe), with minimum effort, reasonable model credibility and extraordinary computational efficiency.

The HE is constructed based on the version of the “*abcd*” model with incorporation of a snow component (Thomos, 1981; Martinez and Gupta, 2010), and several modifications have been made to improve its applicability for practical global applications. First, in order to minimize the number of parameters for ease of use, we set the values for two of the snow-related parameters (i.e., temperature threshold above or below which all precipitation falls as rainfall or snow) from literature (Wen et al., 2013) and only kept one tunable snow-related parameter m – snow melt coefficient ($0 < m < 1$) instead of involving three parameters in the calibration process. Second, we introduce the baseflow index (BFI) into the calibration process to improve the partition of total runoff between the direct runoff and baseflow. Third, other than the lumped scheme as previous studies used, we also explore the values of model application in distributed scheme with a grid resolution of 0.5 degree. The detailed model equations and model parameters can also be found in Appendix A and Table S1 of Liu et al. (2017).

In general, the distributed and lumped schemes of the HE have comparably good capability in simulating spatial and temporal variations of the water balance components (e.g., total runoff, direct runoff, baseflow, evapotranspiration). Meanwhile, the distributed scheme has slightly better performance than the lumped one (e.g., capturing spatial heterogeneity), and also provides grid-level estimates that the lumped scheme does not provide. Additionally, the distributed scheme performs better in extreme climate regimes (e.g., Arctic, North Africa) and Europe. However, the distributed scheme incurs two orders of magnitudes higher computation cost as compared to the lumped scheme. Therefore, the lumped scheme could be an appropriate HE – reasonable predictability and high computational efficiency. At the same time, the distributed scheme could be a suitable alternative for research questions that hinge on grid-level spatial heterogeneity. In terms of recommendations for using the HE, the users are referred to Section 3.4 in Liu et al. (2017).

2. Structure of the HE

We include both a lumped and distributed scheme in the HE for the user's choice. Both schemes are implemented in a monthly time step. In the lumped scheme, each of the global 235 river basins is lumped as a single unit, and each of the climate inputs represent the lumped average across the entire basin, and thus all the model outputs are lumped as well. In terms of the distributed one, each 0.5-degree grid cell has its own input data, and likewise, the model outputs are simulated at the grid-level. Although the two schemes differ in the spatial resolution of their inputs and outputs, their within-basin parameters are uniform. Note that lateral flows between grid cells and basins are not included at this stage for the HE.

3. Emulation by using the HE

The emulation of GHMs by using the HE is mainly achieved by calibrating and validating the HE against the target GHMs. To improve the accuracy of the simulated total runoff and the partition between direct runoff and baseflow, the baseflow index (BFI) is introduced into the objective function during the calibration process. On one side, we maximize Kling-Gupta efficiency (KGE) (Gupta et al., 2009), which is used as a metric to measure the accuracy of the simulated total runoff relative to the benchmark runoff. On the other side, we also nudge the simulated BFI towards the benchmark BFI (here we treat the benchmark BFI as the observed) – the mean BFI of the four products from (Beck et al., 2013). We then conduct parameter optimization by utilizing a Genetic Algorithm (GA) routine (Deb et al., 2002). Details on the objective function and the equations used in the model calibration process can be found in Section 2.4 of Liu et al. (2017).

It is worthwhile to mention that we use the variable infiltration capacity (VIC) model as an example to illustrate the capability of the HE in emulating GHMs in this version of HE documented here, and thus the runoff product from the VIC model is used as the benchmark runoff that the HE is calibrated and validated against. However, the use of the HE is not tied to the VIC model, users can utilize the framework of the HE with any alternative input climate data and benchmark data of water budgets (e.g., runoff, evapotranspiration (ET)), and recalibrate and revalidate the HE to emulate other complex GHMs of interest, to meet their own needs. For example, if one user wants to emulate the ET estimates of the Community Land Model (CLM), the user can use the CLM ET product (or both the ET and runoff products) as the benchmark product and replace

runoff with ET in the objective function (or add ET into the objective function by employing multi-objective approach) to achieve their goal.

4. Main source codes

4.1 Codes tree of main programs

The main codes include the calibration and validation of the HE as detailed below. The two schemes can be run independently based on user's needs. The validation relies on the calibration and it takes the calibrated parameters from the calibration process as inputs for the validation. The validation evaluates the effectiveness of the HE in reproducing the target GHM being emulated.

4.1.1 Calibration process

**Table 1: The code tree of main programs for calibration
(calibration/lump/ or calibration/dist/ directory)**

Source code file name	Role of each code file
Main_HE_cal_lump.m (or Main_HE_cal_dist.m)	Main program that calls other functions for calibration, evaluation and saving variables
abcd.m	Function for calculating water budgets, and it takes input data and yields output of water fluxes and pools.
snowpartition.m	Function for partitioning precipitation between rainfall and snow
ObjFun_abcd_lump.m, (or ObjFun_abcd_dist.m)	Objective function
KGE.m	Function for calculating Kling-Gupta efficiency
Calibration_lump.m (or Calibration_dist.m)	Function for calibrating the HE against the target model being emulated to get basin-specific parameters

The code trees for the HE calibration are shown in Table 1. The main program consists of the main executable (Main_HE_cal.m), water budget calculation (abcd.m), snow partition (snowpartition.m), objective function (ObjFun_abcd_lump.m or ObjFun_abcd_dist.m), Kling-Gupta efficiency calculation (KGE.m), and implementation of calibration (Calibration_lump.m or Calibration_dist.m). The main executable file (Main_HE_cal.m) calls other functions.

4.1.2 Validation process

The code trees of the HE validation are shown in Table 2. Similar to the calibration process, the main program for validation consists of the main executable (Main_HE_val.m) that calls other functions, water budget calculation (abcd.m), snow partition (snowpartition.m), Kling-Gupta efficiency calculation (KGE.m), and implementation of validation (Validation_lump.m or Validation_dist.m).

**Table 2: The code tree of main programs for validation
(validation/lump/ or validation/dist/ directory)**

Source code file name	Role of each code file
Main_HE_val_lump.m (or Main_HE_val_dist.m)	Main program that calls other functions for validation evaluation and saving variables
abcd.m	Function for calculating water budgets, and it takes input data and yields output of water fluxes and pools.
snowpartition.m	Function for partitioning precipitation between rainfall and snow
KGE.m	Function for calculating Kling-Gupta efficiency
Validation_lump.m (or Validation_dist.m)	Function for obtaining calibrated parameters from output of the calibration process and validating the HE against the target model being emulated

4.2 Codes and data availability

The hydrological emulator (HE) is freely available on GitHub (<https://github.com/JGCRI/hydro-emulator/>). We have released the version of the specific HE v1.0.0 referenced in this paper on <https://github.com/JGCRI/hydro-emulator/releases/tag/v1.0.0>, where the user's manual, source code (written in Matlab), all related inputs, calibrated parameters and outputs for each of the global 235 basins, as well as the detailed Readme file are available. In addition, the HE documented here has been translated into Python and is being incorporated into Xanthos (Li et al., 2017), which is an open-source global hydrologic model that allows users to run different combinations of evapotranspiration, runoff, and routing models. The HE will be the default runoff model used in Xanthos 2.0 and will be available on GitHub (<https://github.com/JGCRI/xanthos>).

5. Input data

5.1 Climate data

The climate data needed for the HE only involve monthly precipitation, monthly mean, maximum and minimum air temperature (Table 3). The climate data used in the example of emulating the VIC model, which is also available on Github (<https://github.com/JGCRI/hydro-emulator/>), is obtained from WATCH (Weedon et al., 2011), spanning the period of 1971-1990, and it is 0.5-degree gridded global monthly data. The climate data is used for model simulation over the global 235 major river basins (Kim et al., 2016). Additionally, we use the Hargreaves-Samani method (Hargreaves and Samani, 1982) to estimate potential evapotranspiration (PET), which is a required input

for the HE, which needs mean, maximum and minimum temperature climate data for the calculation. Note that the use of the HE does not hinge on the WATCH data, the users can use any alternative climate input data to drive the HE to emulate GHMs of interest.

Table 3: Required climate input for the HE

Climate variables	Time step	Spatial resolution	Sources
Precipitation	Monthly	0.5-degree	WATCH (Weedon et al., 2011),
Mean air temperature	Monthly	0.5-degree	
Maximum air temperature	Monthly	0.5-degree	
Minimum air temperature	Monthly	0.5-degree	

5.2 Benchmark runoff product

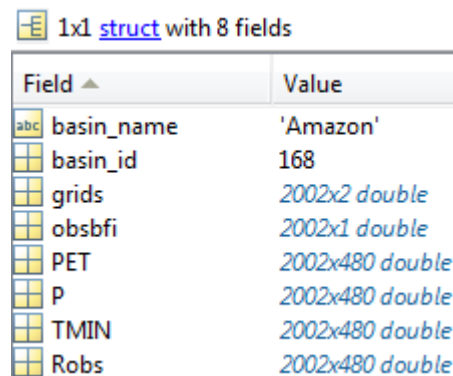
Based on the GHM that the HE is used to emulate, runoff product from that specific GHM is required for the emulation process. In the example documented in Liu et al. (2017) and the codes/data available on Github (<https://github.com/JGCRI/hydro-emulator/>), we use the simulated VIC runoff as the benchmark product that the HE is calibrated and validated against because the VIC is used as the target GHM being emulated in the example. The benchmark product needs to be converted to monthly time step and 0.5-degree if it is in a finer scale, so that the spatial and temporal scales of the benchmark will be consistent with the output from the HE. The VIC runoff product used here is a global simulation with a daily time step and spatial resolution of 0.5 degree for the period of 1971-2010, and the VIC daily runoff is aggregated to monthly data to be consistent with the temporal scale of the HE.

5.3 Baseflow Index (BFI)

BFI is needed as it is incorporated into the objective function to improve the partition of total runoff between direct runoff and baseflow. Global grid-specific BFI data is derived from the mean of the four BFI products in Beck et al. (2013).

5.4 Input data organization

All the input data, include climate data, benchmark runoff and the baseflow index are organized at basin-level, so that the simulation for each of the global 235 basins can be conducted separately, and the users can select basins of their interest. The delineation of the global 235 basins follows that used in Kim et al. (2016). As documented in Section 2, in the lumped scheme each of the data input is the lumped average across the entire basin for each of the global 235 river basins, as opposed to each 0.5-degree grid cell has its own data inputs in the distributed one. Basin-specific input data for the global 235 river basins are available at the directory of WATCH_basin_data, and they are saved in Matlab structure array format with every variable has their own values (.mat). Note that the calculated PET is provided in the input data rather than the mean, minimum and maximum temperature because the HE requires the PET as input rather than temperature.



Field ▲	Value
basin_name	'Amazon'
basin_id	168
grids	2002x2 double
obsbfi	2002x1 double
PET	2002x480 double
P	2002x480 double
TMIN	2002x480 double
Robs	2002x480 double

Figure 1. Input data for the Amazon basin (Matlab structure array)

Figure 1 presents the example of input data in the structure array “WATCH_basin_grid” for the Amazon basin, which includes 8 variables listed in Table 4. Specifically, the “basin name” is Amazon, “basin_ID” is 168 (the basin ID here is not universal and may be different from other data sources), “grids” represents the longitude and latitude of the 2002 grid cells within the Amazon basin, and “BFI” presents grid-specific baseflow index (BFI) derived from the mean of the four BFI products in Beck et al. (2013). In addition, there are 4 other variables “PET”, “P”, “TMIN” and “Robs” with grid- and month-specific values, and they are saved in two-dimension matrix (2002 × 480), the 1st dimension stands for the 2002 within-basin grid cells, and the 2nd dimension represents the 480 months from 1971-2010. “PET” is potential evapotranspiration (PET) calculated from the Hargreaves-Samani method (Hargreaves and Samani, 1982) with mean, maximum and minimum temperatures from the WATCH data (Weedon et al., 2011). “P” and “TMIN” represent precipitation and minimum air temperature, and “Robs”

Table 4 Descriptions for the input data

Variable	Standing for	Comments
basin_name	basin's name	Use that in Kim et al. (2016)
basin_id	basin's ID number	Not for universal use, may be different from that of other data sources
grids	Longitude and latitude of all within-basin grid cells	Grid-specific
obsbfi	Observed baseflow index	Grid-specific
PET	Potential evapotranspiration	Grid- and month-specific
P	Precipitation	Grid- and month-specific
TMIN	Minimum air temperature	Grid- and month-specific
Robs	Observed runoff (benchmark)	Grid- and month-specific

stands for observed runoff (in the documented example we treat the benchmark VIC runoff product as the observed runoff).

6. Output data

6.1 Differences in the outputs from calibration and validation

The output data for the calibration and the validation process is a bit different, and the output for the validation hinges on the output of calibrated parameters from the calibration process. For the calibration process, it provides outputs for the basin-specific calibrated parameters, simulated fluxes (e.g., direct runoff, baseflow, ET, snowmelt) and pools (soil moisture, groundwater, snowpack), and the Kling-Gupta Efficiency (KGE) value of the HE for the calibration period (KGE is used as a metric to measure the performance of the HE, the closer to the value of 1, the better performance the HE presents). With regard to the validation process, it uses the calibrated parameters from the calibration process as input, and it provides similar outputs as that of the calibration process except for the calibrated parameters.

6.2 Output data organization

Similar to the input data, all the output data is saved in the format of Matlab structure array. Figure 2 shows the example of output data in the structure array “dist_cal” for the Amazon basin with use of the distributed scheme for the calibration period (1971-1990). It includes 21 variables (Figure 2, Table 5), with input data precipitation, PET, observed runoff (Robs) and observed BFI (obsbfi) remaining in the array, this is for the user's convenience so that they can easily compare the simulated values to the observed ones with use of the same data array. Variables about basin information, the basin ID

(basin_id), basin name (basin_name) and longitude/latitude of within-basin grid cells (grids) are also kept in the output. The variable “pars” shows the calibrated parameter values for the 5 parameters (a,b,c,d,m) in the HE for the Amazon basin. Similar as the input data format, other variables on water fluxes and pools with grid- and month-specific values are saved as two-dimension matrix (2002×240) in the structure array “dist_cal”, the 1st dimension stands for the 2002 within-basin grid cells, and the 2nd dimension represents the 240 months for the calibration period from 1971-1990. Other than the variables listed in Table 4, the output data also include other 14 variables listed

1x1 struct with 20 fields

Field ▲	Value
basin_name	'Amazon'
basin_id	168
pars	[0.9241,0.9996,0.6512,0.7914,0.0039]
KGE	0.9788
P	2002x240 double
PET	2002x240 double
Rsim	2002x240 double
Robs	2002x240 double
ET	2002x240 double
GW	2002x240 double
SM	2002x240 double
Recharge	2002x240 double
DR	2002x240 double
baseflow	2002x240 double
snowpack	2002x240 double
Rainfall	2002x240 double
snowfall	2002x240 double
snowmelt	2002x240 double
obsbfi	2002x1 double
simbfi	2002x1 double

Figure 2. Output data for the Amazon basin from the calibration process (Matlab structure array)

in Table 5. Note that the variable name in the output structure array maybe a bit different than that in the codes of the HE (abcd.m), for example, the snowpack is denoted as “XS” in the source codes but in the output data it is denoted as “snowpack”. This is because the codes try to follow the convention of denotations in literature, but the output tries to make the variable names as intelligible as possible for the user's convenience.

Table 5 Descriptions for the output data

Variable	Standing for	Comments
pars	Parameters	The 5 parameters (a,b,c,d,m) in the HE, it is assumed uniform across a basin
KGE	King-Gupta Efficiency	Used to measure the HE's performance
Rsim	Simulated runoff	Grid- and month-specific
ET	Evapotranspiration	Grid- and month-specific
GW	Groundwater	Grid- and month-specific
SM	Soil moisture	Grid- and month-specific
Recharge	Recharge to groundwater	Grid- and month-specific
DR	Direct runoff	Grid- and month-specific
Baseflow	Baseflow from groundwater discharge	Grid- and month-specific
snowpack	Snowpack accumulation	Grid- and month-specific
Rainfall	Rainfall partitioned from precipitation	Grid- and month-specific
snowfall	Snowfall partitioned from precipitation	Grid- and month-specific
snowmelt	Snowmelt	Grid- and month-specific
simbfi	Simulated baseflow index	Grid -specific

7. How to run the HE

The HE documented here is written in Matlab, and it can be run either through the Matlab software or via command line. Note that in either way the working directory needs to be navigated to the location where the main program codes are located (see Table 1 and 2) before running the HE. After downloading all the related codes/data from the Github and navigating the working directory to the right location, below are the two ways:

- 1) If run the HE in the Matlab software, first open the main program file for the calibration (e.g., Main_HE_cal_dist.m) or validation process (e.g., Main_HE_val_dist.m) that needs to be run, and then just click the symbol “run” to run it.
- 2) If run the HE via command line, execute the following command:

```
matlab -r “run('main program file')”
```

For example, if run the main program for the calibration process using the distributed scheme (Main_HE_cal_dist.m), the command will be:

```
matlab -r “run('Main_HE_cal_dist.m')”
```

After the running is finished through either way, all related output data will be saved in the current working directory.

Reference

- Beck, H.E. et al., 2013. Global patterns in base flow index and recession based on streamflow observations from 3394 catchments. *Water Resour. Res.*, 49(12): 7843-7863.
- Deb, K., Pratap, A., Agarwal, S., Meyarivan, T., 2002. A fast and elitist multiobjective genetic algorithm: NSGA-II. *IEEE transactions on evolutionary computation*, 6(2): 182-197.
- Gupta, H.V., Kling, H., Yilmaz, K.K., Martinez, G.F., 2009. Decomposition of the mean squared error and NSE performance criteria: Implications for improving hydrological modelling. *J. Hydrol.*, 377(1): 80-91.
- Hargreaves, G.H., Samani, Z.A., 1982. Estimating potential evapotranspiration. *Journal of the Irrigation and Drainage Division*, 108(3): 225-230.
- Kim, S.H. et al., 2016. Balancing global water availability and use at basin scale in an integrated assessment model. *Clim. Change*, 136(2): 217-231.
- Li, X., Vernon, C.R., Hejazi, M.I., Link, R.P., Feng, L., Liu, Y., Rauchenstein, L.T., 2017, Xanthos – A Global Hydrologic Model, *Journal of Open Research Software*, 5(1), p.21.
- Liu, Y., Hejazi, M.A., Li, H., Zhang, X., (2017), A Hydrological Emulator for Global Applications, *Geoscientific Model Development Discussions*, DOI: 10.5194/gmd-2017-113
- Thomas, H., 1981. Improved methods for national water assessment. Report WR15249270, US Water Resource Council, Washington, DC.
- Martinez, G.F., Gupta, H.V., 2010. Toward improved identification of hydrological models: A diagnostic evaluation of the “abcd” monthly water balance model for the conterminous United States. *Water Resour. Res.*, 46(8).
- Weedon, G. et al., 2011. Creation of the WATCH forcing data and its use to assess global and regional reference crop evaporation over land during the twentieth century. *J. Hydrometeorol.*, 12(5): 823-848.
- Wen, L., Nagabhatla, N., Lü, S., Wang, S.-Y., 2013. Impact of rain snow threshold temperature on snow depth simulation in land surface and regional atmospheric models. *Adv. Atmos. Sci.*, 30(5): 1449-1460.