PROCLIAS

Process-based models for climate
impact attribution across sectors

# “Using ISIMIP model output with ILAMB and ESMValTool. First insights in application / modification and challenges towards its usability as a quality assessment tool for impact models”

Technical documentation By
Emmanuel Nyenah, Hannes Müller Schmied & Aristeidis Koutroulis

Created within Taskgroup 1.2 within a virtual mobility grant.

*EU Cost-Action numb*er: CA19139
*Date*: 13/9/2021

cost
EUROPEAN COOPERATION
IN SCIENCE & TECHNOLOGY

This report is based upon work from COST Action PROCLIAS (CA19139), supported by COST European Cooperation in Science & Technology

COST is supported
by the Horizon 2020
Framework Programme
of the European Union

# Table of Contents

1 Summary ....2

2 ILAMB ....3

2.1 Overview ....3

2.2 Installation ....3

2.3 Data basis ....4

2.3.1 Observation data ....4

2.3.2 Simulation data ....4

2.4 Assessment ....5

2.4.1 Set-up ....5

2.4.2 Modifications of ILAMB ....6

2.5 Results ....6

2.5.1 TWSA ....6

2.5.2 GRDC ....12

2.6 Pro's and Con's (caveats) ....15

3 ESMValTool ....16

3.1 Overview ....16

3.2 Installation ....17

3.3 Pro's and Con's (caveats) ....17

4 Conclusion & Outlook ....17

Acknowledgements ....18

References ....18

Appendix ....20

# 1 Summary

The Inter-Sectoral Impact Model Intercomparison Project (ISIMIP) provides a framework for the intercomparison of global and regional-scale impact models within and across several sectors and to enable coordinated multi-sectoral assessments of different risks and their aggregated effects (Rosenzweig et al., 2017; Schellnhuber et al., 2014). It is of importance to ensure plausible model outputs, which can be achieved by applying quality control (QC) and quality assessment (QA) tools. Quality control (QC) tools are meant to proof consistency of metadata, plausible ranges and file structure, whereas quality assessment (QA) tools are designed to compare model output to reference data. Comparing models against reference data is an integral part of model development and can be understood as prerequisite for impact studies in terms of judging the reliability of those models to realistically replicate earth system processes (Krysanova et al., 2020). Adopting existing QA/QC tools to an automatic framework for the models used in ISIMIP could support modelling activities and coordination between modelling teams through the umbrella of PROCLIAS. Whereas a QC tool is already existing and in operation for ISIMIP, the focus of this report is to elaborate first steps towards a QA tool.

An initial paper/website-based review of existing tools, in the frame of the EU Cost Action CA19139 PROCLIAS Taskgroup 1.2., entitled with "Automatic QC/QA for impact model output", showed a range of tools (e.g., ILAMB(https://www.ilamb.org/) and ESMValTool (https://www.esmvaltool.org/)) which are in routine use e.g., within the Coupled Model Intercomparison Project (CMIP) framework.

This report documents the results of a VM Grant embedded within PROCLIAS TG1.2 with the aim to apply real ISIMIP model output data from the global water sector (as a pilot sector) to the pre-selected quality assessment tools (ILAMB and ESMValTool). By that, the usability of those tools for impact model evaluation is tested and a pathway towards adapting those tools for usability within the global water sector and furthermore across the ISIMIP sectors is generated. This project was conducted within a 2-month virtual mobility grant acquired by the first author of this report.

After investigation of the ILAMB and ESMValTool, we do see value for using ILAMB as a potential quality assessment tool for model evaluation within the ISIMIP framework. ILAMB has been modified by the grant holder to include sector-specific metrics such as KGE and NSE in addition to the default metrics for benchmarking analysis. Also, this tool has a feature which allows for model evaluation with defined domain of interest but will require further code development for basin specific benchmarking. It is key to note that it was technical impossible in the framework of this grant to transform the ISIMIP model output to be compliant with the ESMValTool.

ILAMB was tested for two variables, Terrestrial Water Storage (comparing WaterGAP with GRACE data) and observed streamflow data (comparing WaterGAP and PCR-GLOBWB with GRDC data). Those results show the capability of the tool to investigate Individual model performance as well as intercomparison between impact models in a consistent framework.

Currently, sector specific metric added to ILAMB show results in the console rather than in the result webpage. Future work could improve this caveat. Also, a CMORization tool which can make ISMIP data compliant with ESMValTool should be explored to enable the use of ESMValTool for benchmark analysis. Finally in addition to this above stated tools other potential QA tools could also be explored for model evaluation within this sector.

# 2 ILAMB

## 2.1 Overview

ILAMB is an open-source model benchmarking software package that generates graphical diagnostics and scores land model performance (Collier et al., 2018). It assesses model performance for variables in categories of biogeochemistry, hydrology, radiation and energy and climate forcing. For each of these variables, the packages generate graphical diagnostics and score model performance for the period mean over whole years and its bias, bias score, RMSE, RMSE score, spatial distribution, interannual coefficient of variation, and seasonal cycle and long-term trend (Collier et al., 2018). ILAMB also supports benchmark analysis per region. There are already defined customed regions in ILAMB which can be found at https://www.ilamb.org/doc/ilamb_run.html. For benchmarking analysis using user defined regions, more information can be found at https://www.ilamb.org/doc/custom_regions.html.

A simplified working framework of ILAMB is shown in Figure 1.

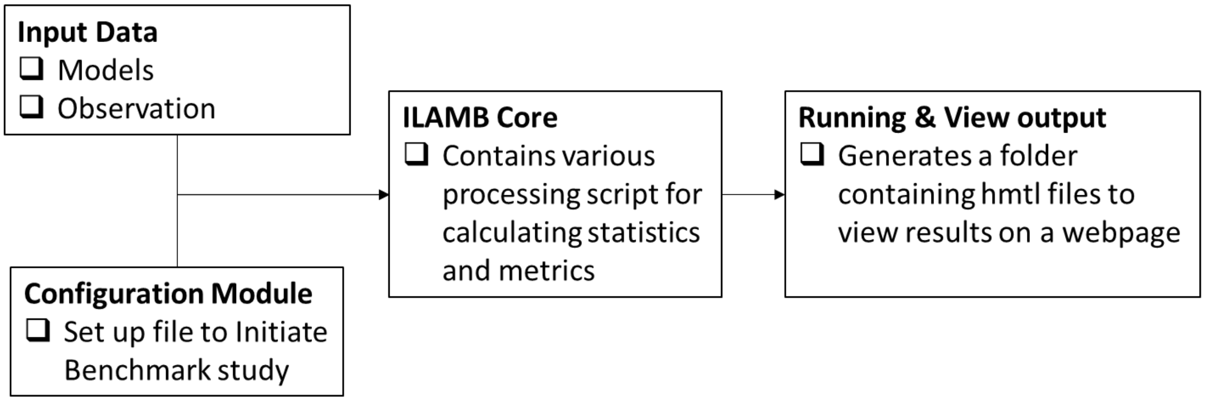

Input Data
❑ Models
❑ Observation
ILAMB Core
❑ Contains various processing script for calculating statistics and metrics
Running & View output
❑ Generates a folder containing hmtl files to view results on a webpage
Configuration Module
❑ Set up file to Initiate Benchmark study

*Figure 1: Simplified ILAMB framework.*

## 2.2 Installation

The process described below works on running ILAMB on HPC clusters with programming in bash or linux environment.

1. Download MiniConda which is a python environment via:
   ```
   $ wget https://repo.continuum.io/miniconda/Miniconda3-latest-Linux-x86_64.sh
   ```
2. Run the installation script:
   ```
   $ bash Miniconda3-latest-Linux-x86_64.sh -b -p $VSC_DATA/miniconda
   ```
3. Once the python environment is ready, run the command $ conda install ILAMB, to get the latest software from the ILAMB repository.

For encountered errors during installation, a very detailed procedure can be found on the ILAMB documentation webpage https://www.ilamb.org/doc/index.html.

## 2.3 Data basis

### 2.3.1 Observation data

#### *2.3.1.1 Gravity Recovery and Climate Experiment (GRACE)*

Total water storage anomaly (TWSA) obtained from GRACE Tellus CSR observations are already available in ILAMB benchmark repository. These reference data have a spatial resolution of 0.5° x 0.5° and a monthly temporal resolution (Landerer & Swenson, 2012). The data retrieved spans the period between 2002-2014.

#### *2.3.1.2 Global Runoff Data Centre (GRDC)*

Monthly observed river discharge was obtained for 12 GRDC stations, see Table 1, (GRDC, 2021). This retrieved data spans the period between 2001-2010. The rationale of selecting those 12 stations is to test the usability of ILAMB with newly added benchmark data. To make GRDC data ILAMB compliant, Matthias Büchner (PIK Potsdam) developed a python code for such conversion. This code is hosted on https://github.com/ISI-MIP/GRCD_convert & https://github.com/ISI-MIP/GRCD_convert/issues.

*Table 1: GRDC stations for benchmark analysis.*

| Station ID | River Name | Station Name |
|---|---|---|
| 1147010 | CONGO RIVER | Kinshasa |
| 1159100 | ORANGE RIVER | Vioolsdrif (27811003) |
| 2181900 | YANGTZE RIVER | Datong |
| 2569003 | MEKONG RIVER | Kompong Cham |
| 2903430 | LENA | Stolb |
| 3265601 | PARANA, RIO | Timbues |
| 3629000 | AMAZON RIVER | Obidos - Porto |
| 4127800 | MISSISSIPPI RIVER | Vicksburg, Ms |
| 4208025 | MACKENZIE RIVER | Arctic Red River |
| 5204268 | MURRAY RIVER | Lock 9 Upstream (764.8 Km) |
| 6742900 | DANUBE RIVER | Ceatal Izmail |
| 6977100 | VOLGA | Volgograd Power Plant |

### 2.3.2 Simulation data

The selection of model outputs from the global water sector used in this study was solely arbitrary and only for test purposes. Both models used have a daily temporal resolution and a spatial resolution of is 0.5° x 0.5° (~ 55 km by 55 km at the equator).

#### *2.3.2.1 WaterGAP*

The global freshwater use and availability model WaterGAP2 calculates water use for five sectors (irrigation, domestic, manufacturing, cooling water for electricity generation, and livestock) that are then processed by the Groundwater Surface Water USE (GWSWUSE) submodule to quantify both net water abstractions from surface water and from groundwater resources (Müller Schmied et al., 2021). The WaterGAP Global Hydrology Model (WGHM) considers net water abstractions to calculate changes in water storage compartments as well as water flows between these compartments based on water balance equations, including groundwater recharge, evapotranspiration, and river discharge (Müller

Schmied et al., 2014, 2021). WaterGAP is calibrated in a basin-specific manner to match long-term annual observed river discharge at 1319 river basins that cover ~54 % of global drainage area (Müller Schmied et al., 2021).

##### 2.3.2.1.1 Total Water Storage

Simulated total water storage (TWS) is modelled by WaterGAP2.2d model with WFDEI climate forcing (Müller Schmied et al., 2020, 2021). TWS from 2002-2014 is retrieved with historical water use and reservoir operation. In order to compare the model with observation from GRACE, monthly anomalies of TWS are computed with a reference mean period from 2004-2009 which is the same reference period as GRACE. GRACE TWSA data exclude Greenland and Antarctica and hence Greenland was masked out from WaterGAP2 which already excludes Antarctica.

##### 2.3.2.1.2 River discharge

Two WaterGAP simulated river discharge output from ISIMIP2a are used in this study. The first time series is produced by driving the WaterGAP2.2 model with Global Soil Wetness Project 3 (GSWP3) climate forcing and the second by the WFDEI climate forcing (Müller Schmied et al., 2020, 2021). These forcings are provided by the Inter-Sectoral Impact Model Intercomparison Project (ISIMIP) in its phase 2a (https://www.isimip.org/about/#simulation-rounds-isimip2a). Daily river discharge data from 2001-2010 is retrieved and aggregated to monthly temporal resolution. Both data sets (WaterGAP2_WFDEI and WaterGAP2_GSWP3) are considering historical water use and reservoir operation.

#### *2.3.2.2 PCR-GLOBWB*

PCR-GLOBWB model calculates the surface water balance and monthly sectoral water demand and incorporates groundwater abstraction at the global scale (Sutanudjaja et al., 2018; Wada et al., 2014).

##### 2.3.2.2.1 River discharge

Simulated PCR-GLOBWB river discharge output from ISIMIP2a with WFDEI climate forcing is used in this study. Daily river discharge data (2001-2010) retrieved also considers historical water use and reservoir operations are used. Discharge data is aggregated to monthly data.

## 2.4 Assessment

### 2.4.1 Set-up

#### *2.4.1.1 Input data*

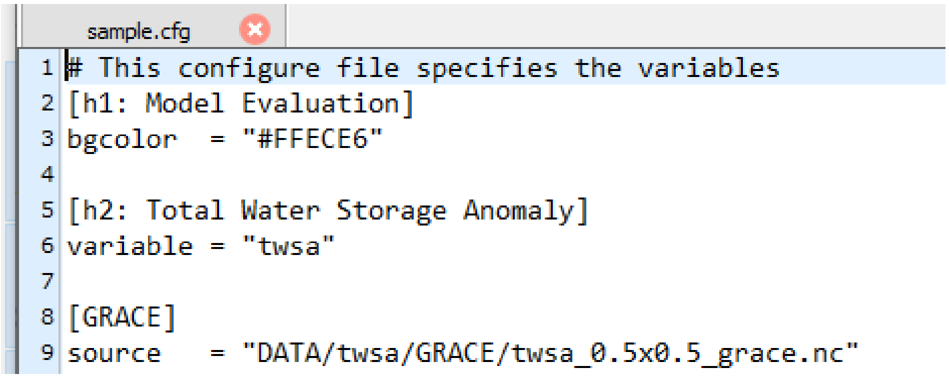
After prior data processing (see appendix A1), benchmark analysis is then performed.

#### *2.4.1.2 Setting up a configuration file*

Figure 2 shows the configuration file example made for TWSA model evaluation. The h1 tag is the top-level heading which denotes benchmarking tasks to be done (Model Evaluation in this case). The hex colour assigned is used to set the background colour of the benchmark results in the result webpage. The h2 tag is a sub level heading which denotes which variable should be analysed (Total Water Storage Anomaly in this case with variable name ‘twsa’).

Finally, the location of the benchmark file is set. The heading [GRACE] denotes a user defined name for the benchmark and the ‘source’ denotes the location where the ILAMB core reads the benchmark file.

More information on setting up a configuration file for benchmark analysis can be found on https://www.ilamb.org/doc/first_steps.html.



```
# This configure file specifies the variables
[h1: Model Evaluation]
bgcolor  = "#FFECE6"

[h2: Total Water Storage Anomaly]
variable = "twsa"

[GRACE]
source   = "DATA/twsa/GRACE/twsa_0.5x0.5_grace.nc"
```

*Figure 2 Configuration file setup.*

### 2.4.2 Modifications of ILAMB

#### *2.4.2.1 Adding Kling–Gupta and Nash–Sutcliffe efficiency metrics*

In addition to the default metrics in ILAMB, Kling–Gupta efficiency, KGE, (Gupta et al., 2009; Knoben et al., 2019) and Nash–Sutcliffe efficiency, NSE, (Nash & Sutcliffe, 1970) are added for benchmark analysis (see appendix A2). By default, KGE and NSE are based on spatially integrated mean (which is the mean time series all GRDC stations used) and the code must be edited otherwise.

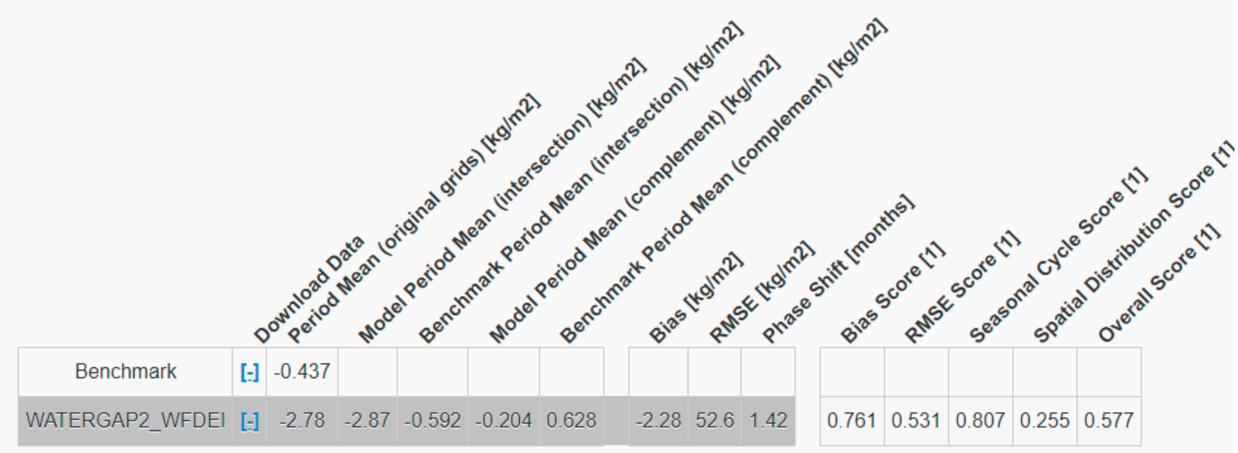
## 2.5 Results

### 2.5.1 TWSA

Figure 3 shows the general performance measures of the comparison WaterGAP2.2d with GRACE. The numbers can be interpreted exemplarily as follows: From Figure 3, WaterGAP2 globally underestimates the mean annual TWSA with a mean bias of -2.3 kg m$^{-2}$ (bias score = 0.76,). Notable regions where WaterGAP2.2d underestimates mean TWSA are Southern Africa, Eastern Australia, northern regions of South America and Western USA (Figure 4). Also, the model tends to majorly overestimate TWSA in regions such Alaska and North Eastern Canada (Figure 4). This might be because WaterGAP2 does not simulate glaciers (Müller Schmied et al., 2021). From the mean annual TWSA (Figure 5), GRACE detects strong decrease of water storage due to glacial mass loss for these regions but WaterGAP2 shows a small increase on average.

An overall good performance based on the RMSE score of 0.53 (RMSE=52.6 kg m$^{-2}$) is shown by the model (Figure 3) but spatial differences are existing (see Figure 6).

From Figure 7, the difference between the maximum of the TWSA annual cycle of the WaterGAP2 and GRACE (phase shift) is also investigated. Global mean phase shift of WaterGAP2 is 1.4 months (seasonal cycle score=0.81) with a strong lag in phase ($-2<\theta<-6$ months) in regions such as Australia, Eastern China, central USA and parts of Northern Africa.



| | Download Data | Period Mean (original grids) [kg/m2] | Model Period Mean (intersection) [kg/m2] | Benchmark Period Mean (intersection) [kg/m2] | Model Period Mean (complement) [kg/m2] | Benchmark Period Mean (complement) [kg/m2] | Bias [kg/m2] | RMSE [kg/m2] | Phase Shift [months] | Bias Score [1] | RMSE Score [1] | Seasonal Cycle Score [1] | Spatial Distribution Score [1 | Overall Score [1] |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Benchmark | [-] | -0.437 | | | | | | | | | | | | |
| WATERGAP2_WFDEI | [-] | -2.78 | -2.87 | -0.592 | -0.204 | 0.628 | -2.28 | 52.6 | 1.42 | 0.761 | 0.531 | 0.807 | 0.255 | 0.577 |

*Figure 3: Summary of global benchmark analysis for TWSA (2002-2014).*

### *2.5.1.1 Spatial Distribution*

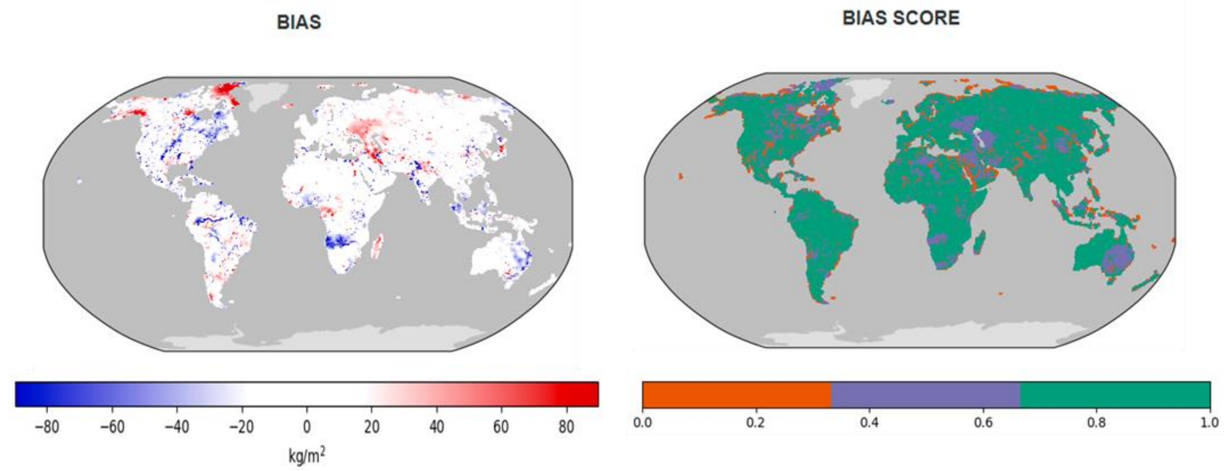


*Figure 4: Annual mean bias and bias score for WaterGAP2_WFDEI_GPCC (2002-2014).*

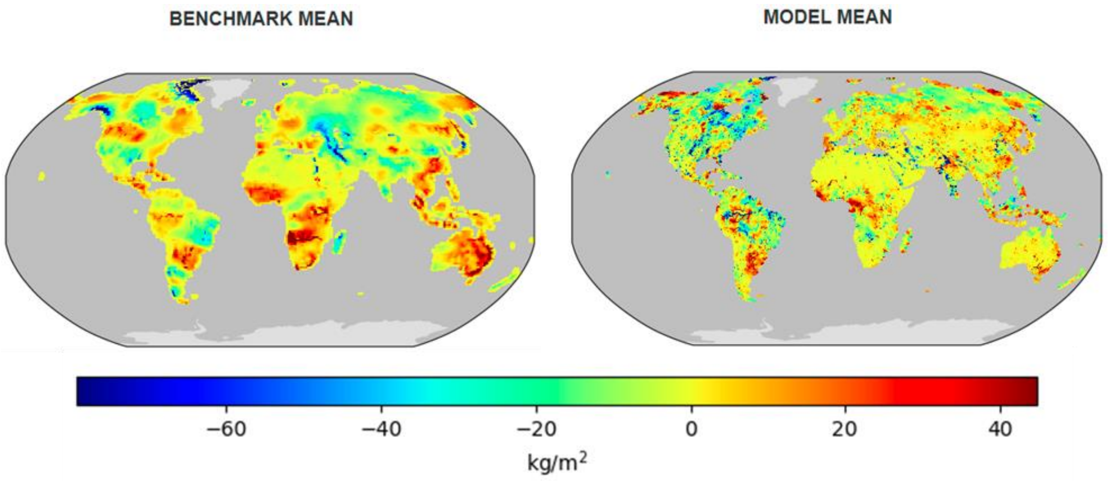


*Figure 5: Annual mean TWSA for GRACE and WaterGAP2_WFDEI_GPCC (2002-2014).*

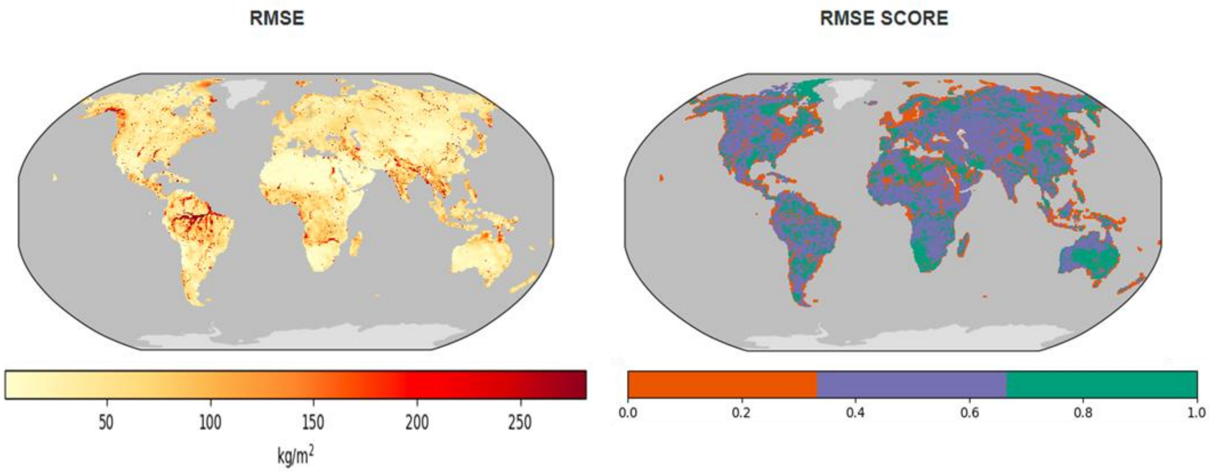


*Figure 6: Annual mean RMSE and RMSE score for WaterGAP2_WFDEI_GPCC (2002-2014).*

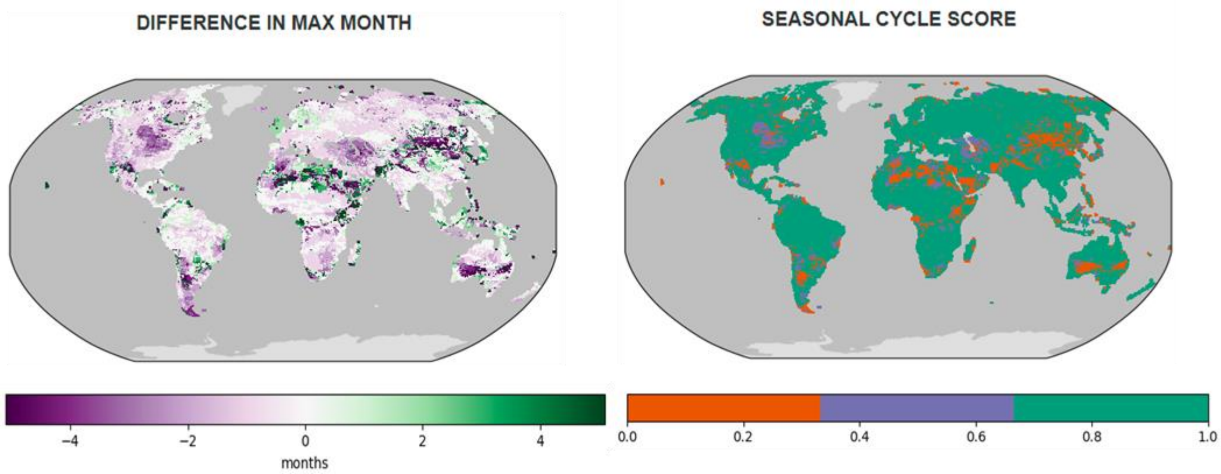


*Figure 7: Annual mean phase shift and score (seasonal cycle score) for WaterGAP2_WFDEI_GPCC (2002-2014).*

### 2.5.1.2 Taylor Diagram

Here the spatial distribution of mean twsa (see Figure 5) between WaterGAP2 and GRACE is visualized by the use of a spatial Taylor diagram. From the spatial Taylor diagram (Figure 8), the deviation of the model (red dot) from the benchmark (black star) can be seen. This deviation is quantified by spatial distribution score ($S_{dist}$=0.255, see Figure 3) which is based on the spatial correlation coefficient and normalised standard deviation of the mean values of model and benchmark (see https://www.ilamb.org/ILAMB_paper.pdf for the theory behind the spatial Taylor diagram).

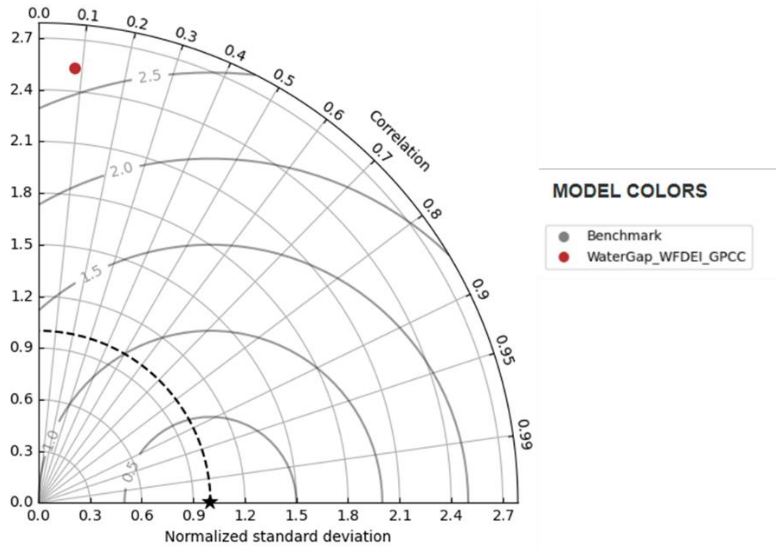


*Figure 8: Taylor diagram for TWSA benchmark analysis for 2002-2014. Benchmark (GRACE) shown in black star.*

### 2.5.1.3 Temporal Distribution

From Figure 9, there is very good agreement between the temporal components of TWSA from GRACE and WaterGAP2. From the annual cycle plot (Figure 9-bottom), it is evident that there is a lag in phase by WaterGAP2. It is important to note that the regional mean title as displayed on the result webpage for the temporal mean timeseries can mean either global or regional depending on the spatial scale of the analysis and hence titles should be changed accordingly.

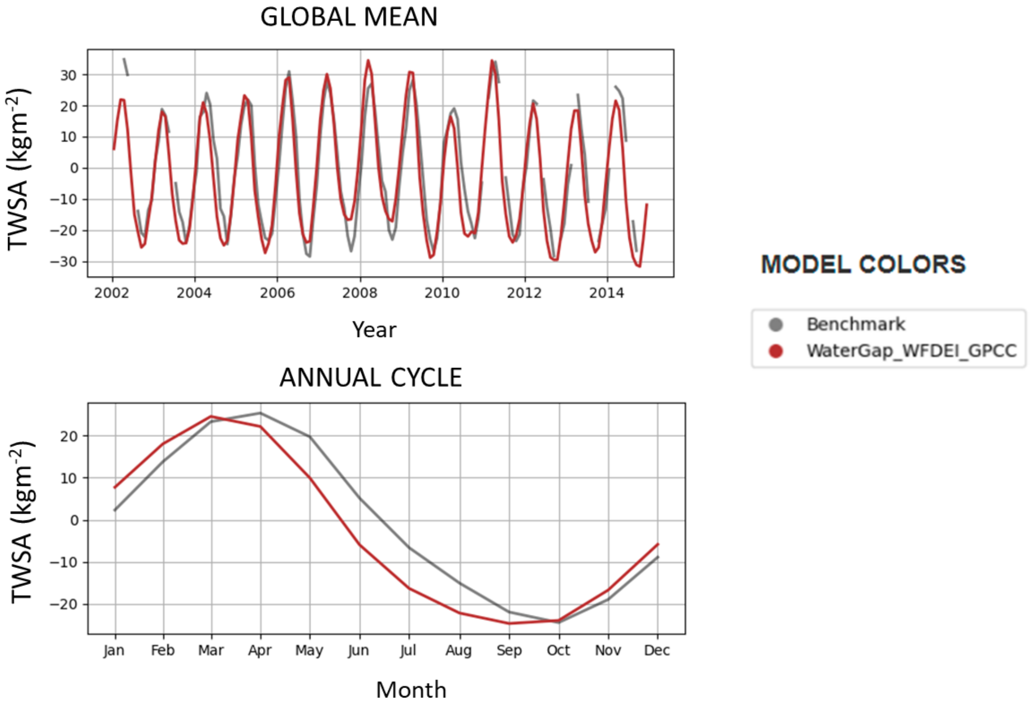
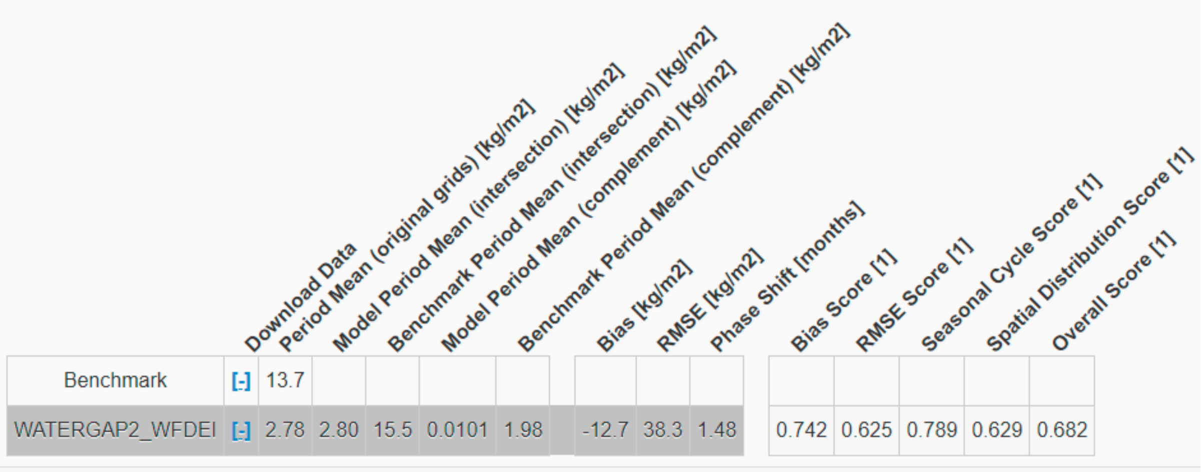


*Figure 9: Global annual mean (top) and annual cycle(bottom) for the period 2002-2014.*

From the overall score of 0.578 which is the weighted sum of all the score in the result summary (Figure 3), WaterGAP2 emerge as a good model in estimating mean TWSA. KGE and NSE obtained for TWSA (the time series of all grid cells globally) are -2.85 and 0.83 respectively.

### *2.5.1.4 Regional Benchmarking (Australia)*

In this section, an interesting function of ILAMB, regional benchmarking is presented. With this function, default regions in ILAMB (defined by shapefiles) can be selected and also, users can perform model evaluation using their own defined region (rectangular selection via bounding coordinates). An exemplary result for regional benchmarking analysis of TWSA for Australia (default region in ILAMB) is explored. We show only the performance summary (Figure 10), RMSE and its score (Figure 11 ), spatially integrated distribution (Figure 12).



| | Download Data | Period Mean (original grids) [kg/m2] | Model Period Mean (intersection) [kg/m2] | Benchmark Period Mean (intersection) [kg/m2] | Model Period Mean (complement) [kg/m2] | Benchmark Period Mean (complement) [kg/m2] | Bias [kg/m2] | RMSE [kg/m2] | Phase Shift [months] | Bias Score [1] | RMSE Score [1] | Seasonal Cycle Score [1] | Spatial Distribution Score [1] | Overall Score [1] |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Benchmark | [-] | 13.7 | | | | | | | | | | | | |
| WATERGAP2_WFDEI | [-] | 2.78 | 2.80 | 15.5 | 0.0101 | 1.98 | -12.7 | 38.3 | 1.48 | 0.742 | 0.625 | 0.789 | 0.629 | 0.682 |

*Figure 10: Summary of Australia benchmark analysis for TWSA (2002-2014).*

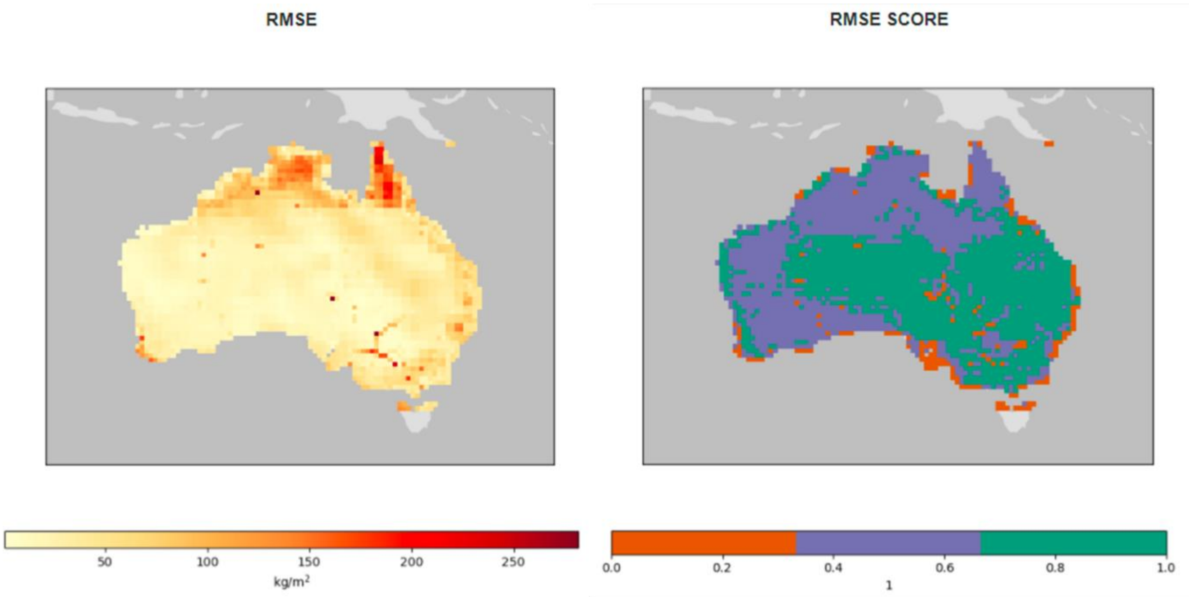


*Figure 11: Annual mean RMSE and RMSE score for WaterGap2_WFDEI_GPCC (2002-2014).*

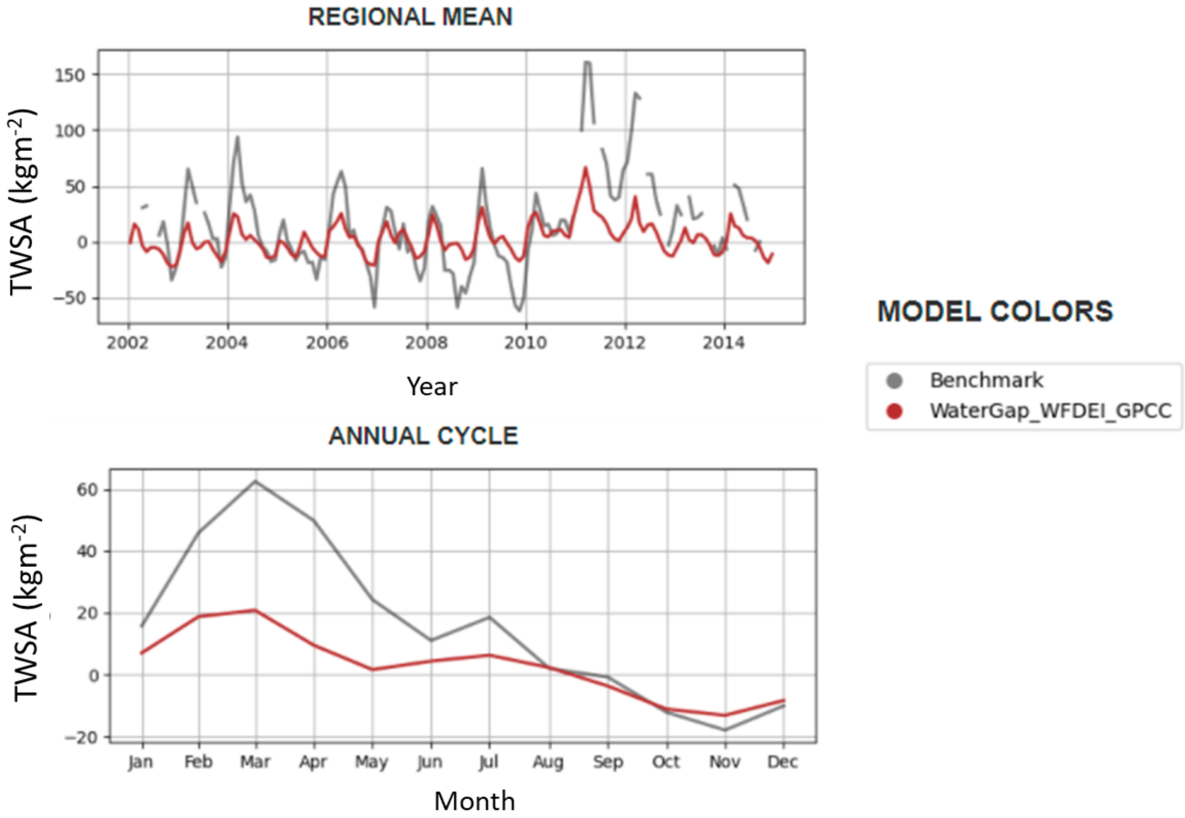


*Figure 12: Annual temporal mean (top) and annual cycle(bottom) for the period 2002-2014.*

## 2.5.2 GRDC

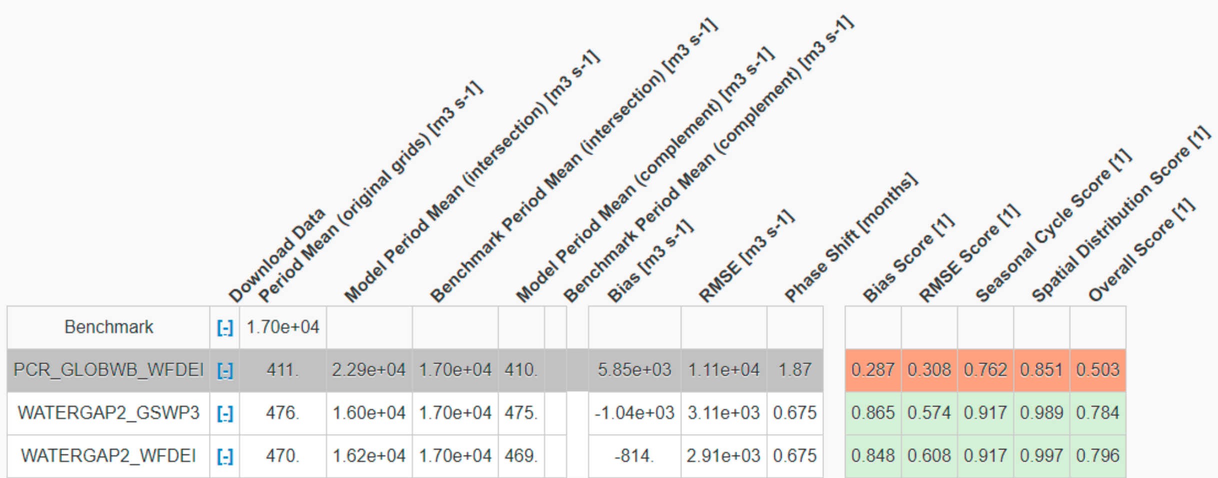
This section explores model performance of WaterGAP2 and PCR-GLOBWB against all 12 GRDC stations used in this study. By default, the metrics are calculated for all benchmark data if not specific regions are defined. From a process-oriented approach this averaging is not meaningful. However, the purpose here is to show how the inclusion of additional reference data and metrics can be used for assessing model output.

The example interpretation of this assessment can be as follows: From Figure 13, PCR_GLOBWB_WFDEI tends to overestimate observed discharge with a bias of 5.9 x $10^3$ $m^3s^{-1}$ (bias score = 0.29) as compared to WaterGAP2 model which tend to underestimate the observed discharge. WaterGap2_GSWP3 has a bias of -1 x $10^3$ $m^3s^{-1}$ (bias score = 0.87) and WaterGAP2_WFDEI with a bias of –8 x $10^2$ $m^3s^{-1}$ (bias score = 0.85).

Based on RMSE (Figure 13), WaterGAP2_WFDEI emerges as the model with the best performance with a RMSE score of 0.608 (RMSE = 2.91 x $10^3$ $m^3s^{-1}$) followed by WaterGAP2_GSWP3 with a score of 0.574 (RMSE = 3.11 x $10^3$ $m^3s^{-1}$). PCR_GLOBWB_WFDEI emerges as the model with least performance with a RMSE score of 0.308 (RMSE = 1.1 x $10^4$ $m^3s^{-1}$).

The phase shift observed for the WaterGAP2 models is 0.68 months whiles that of PCR_GLOBWB_WFDEI has a lag of 1.87 months (Figure 13).

It is important to note that for temporal mean calculation, ILAMB requires initial and final time step of the data used for analysis. If either or both time steps are missing, such grid point (station in the case of GRDC) are omitted. But for spatial mean calculation, ILAMB takes the average of all grid cells and ends up with one overall time series. In the case of GRDC, all stations are considered for the spatial mean.



| | Download Data | Period Mean (original grids) [m3 s-1] | Model Period Mean (intersection) [m3 s-1] | Benchmark Period Mean (intersection) [m3 s-1] | Model Period Mean (complement) [m3 s-1] | Benchmark Period Mean (complement) [m3 s-1] | Bias [m3 s-1] | RMSE [m3 s-1] | Phase Shift [months] | Bias Score [1] | RMSE Score [1] | Seasonal Cycle Score [1] | Spatial Distribution Score [1] | Overall Score [1] |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Benchmark | [-] | 1.70e+04 | | | | | | | | | | | | |
| PCR_GLOBWB_WFDEI | [-] | 411. | 2.29e+04 | 1.70e+04 | 410. | | 5.85e+03 | 1.11e+04 | 1.87 | 0.287 | 0.308 | 0.762 | 0.851 | 0.503 |
| WATERGAP2_GSWP3 | [-] | 476. | 1.60e+04 | 1.70e+04 | 475. | | -1.04e+03 | 3.11e+03 | 0.675 | 0.865 | 0.574 | 0.917 | 0.989 | 0.784 |
| WATERGAP2_WFDEI | [-] | 470. | 1.62e+04 | 1.70e+04 | 469. | | -814. | 2.91e+03 | 0.675 | 0.848 | 0.608 | 0.917 | 0.997 | 0.796 |

*Figure 13: Summary of benchmark analysis for river discharge (2001-2010).*

### 2.5.2.1 Taylor Diagram

The spatial distribution of mean discharge between WaterGAP2, PCRGLOBWB and GRDC is explored here (Figure 14). Both climate forcing of WaterGAP2 (WFDEI and GSWP3) show very small deviation from GRDC

(Blackstar) reflecting in a spatial distribution score of 0.99 for both forcing (see Figure 13). Even though PCR-GLOBWB model deviates from the GRDC, it has a very good spatial distribution score of 0.851 (see Figure 13) .

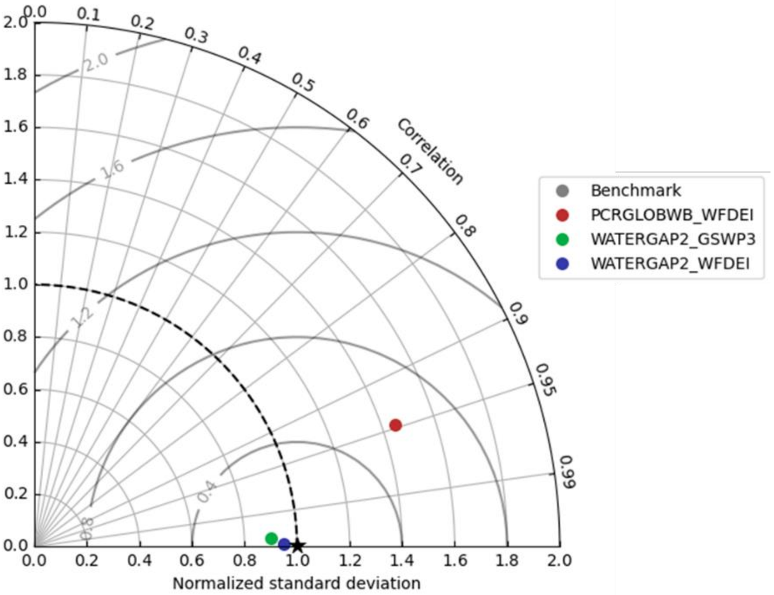


*Figure 14:Taylor diagram for spatial distribution of mean discharge between WaterGAP2 (WFDEI & GSWP3), PCRGLOBWB and GRDC for 2001-2010. Benchmark (GRACE) shown in black star.*

### 2.5.2.2 Temporal distribution

There is very good agreement between the global mean of WaterGAP2 and PCR_GLOBWB_WFDEI with GRDC except that PCR_GLOBWB_WFDEI has very high discharges amplitudes (Figure 15).  Same can be said for the annual cycle (Figure 16). There is also a very good agreement between the monthly anomaly of WaterGAP2 and GRDC while PCR_GLOBWB_WFDEI on average tend to overestimate discharge during June, July and August and underestimate it during the December, January and February (Figure 17).

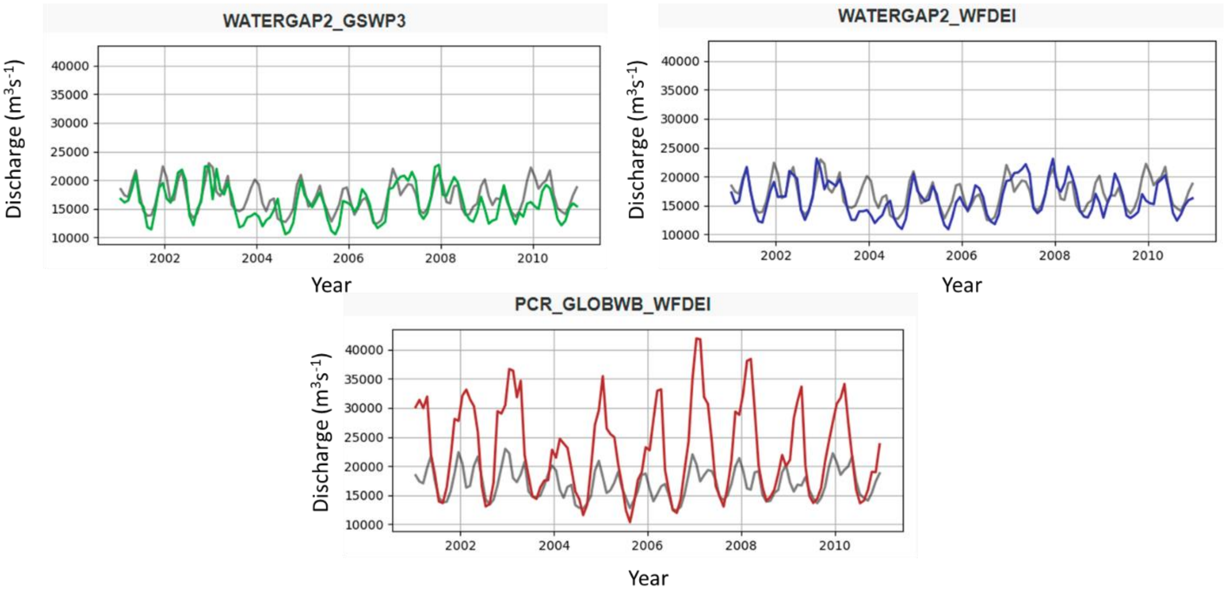


*Figure 15: Spatial average of monthly time series for the 12 basins and for WaterGAP2_GSWP3 (green), WaterGAP2_WFDEI (blue), and PCR_GLOBWB_WFDEI (red) for time period 2001-2010. GRDC is shown in grey.*

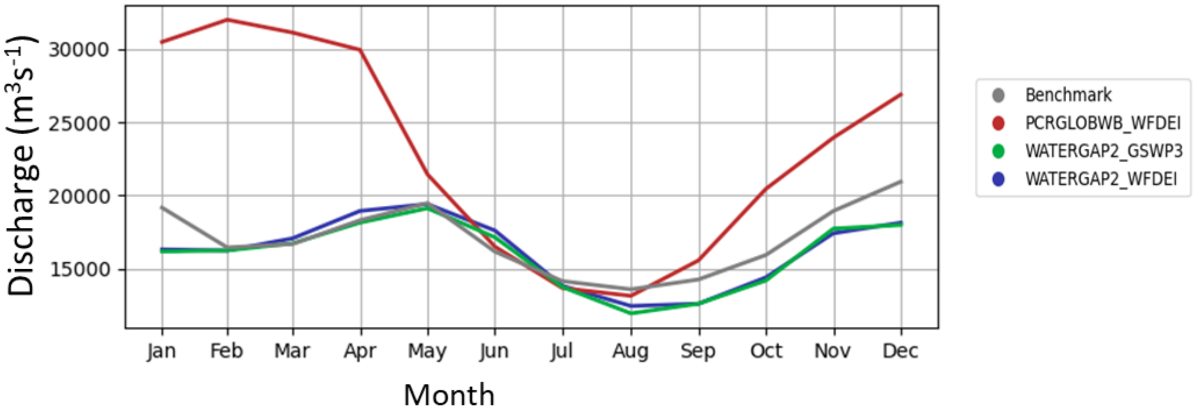


*Figure 16: Spatial average of annual cycle for the 12 basins and the time period 2001-2010.*

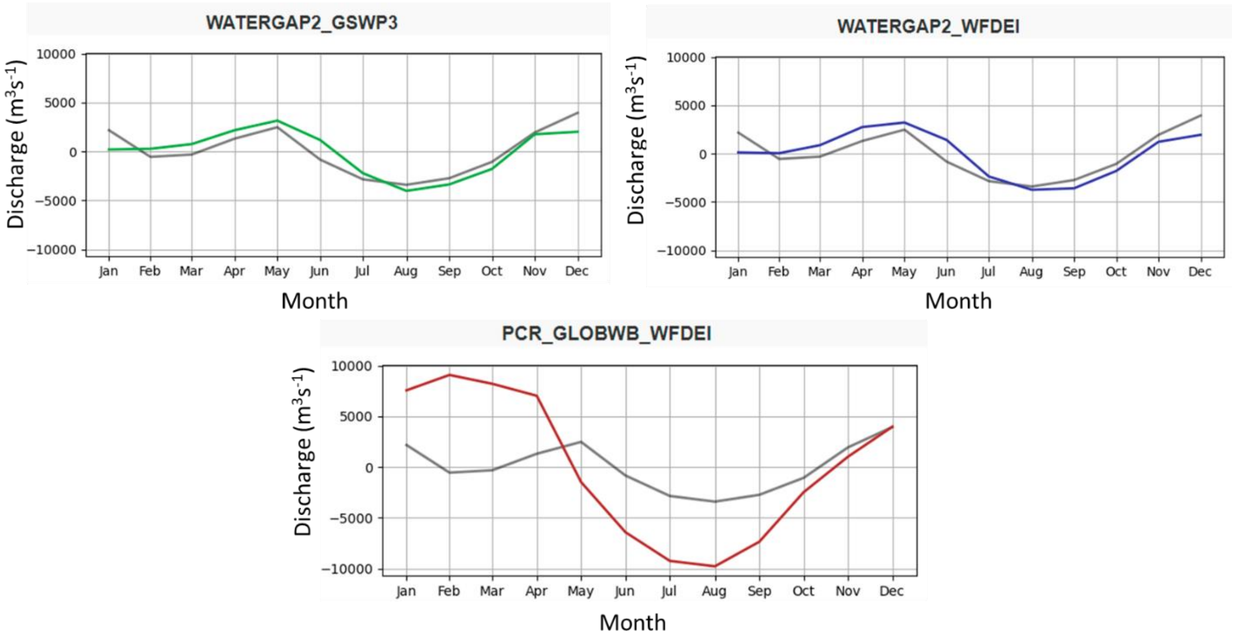


*Figure 17: Spatial average of anomaly for WaterGAP2_GSWP3 (green) , WaterGAP2_WFDEI (blue) , and PCR_GLOBWB_WFDEI (red) for time period 2001-2010. GRDC is shown in grey.*

From Figure 13, WaterGap2_WFDEI emerges as the model with the best performance (overall score=0.796) followed by WaterGap2_GSWP3 (overall score=0.784) and PCR_GLOBWB_WFDEI (overall score=0.503).

### 2.5.2.3 Kling–Gupta efficiency (KGE) & Nash-Sutcliffe efficiency (NSE)

Here we show the results of KGE and NSE for river discharge. There is very good agreement between WaterGAP2_GSWP3, WaterGAP2_WFDEI model and GRDC as reflected by the high score in KGE (see Table 2). PCR_GLOBWB_WFDEI shows less agreement with GRDC. NSE also reflects similar conclusion even though values are lower. Please note that WaterGAP was calibrated with the discharge stations used for this comparison (but for a different time period).

*Table 2: KGE and NSE for river discharge as spatial mean over the 12 GRDC stations.*

| Model | KGE | NSE |
|---|---|---|
| **WaterGap2_GSWP3** | 0.68 | 0.19 |
| **WaterGap2_WFDEI** | 0.7 | 0.28 |
| **PCR_GLOBWB_WFDEI** | -1.1 | -10.3 |

## 2.6 Pro's and Con's (caveats)

An interesting advantage of ILAMB is its ability to perform benchmark analysis for several model output at once, allowing the use of other reference data and its flexibility to add additional metrics.

However, ILAMB is developed for a specific structure of NETCDF input data. Some ISIMIP data follow the ILAMB data structure but have missing time bounds or time interval variables. The time bound is required by ILAMB to precisely match the correct time interval between a confrontation pair (model and benchmark). Information on how to make datasets ILAMB compliant can be found at https://www.ilamb.org/doc/format_data.html.

Data variables should have the same name and units for both observed and model. For example, after calculating the anomaly series for WaterGAP TWS data, its variable name was changed to 'twsa' to enable ILAMB run without errors.

ILAMB allows for benchmarking analysis with defined region or domain of interest. ILAMB provides default regions via shapefile selection but user defined regions which are outside the default ILAMB regions are via rectangular selection. For the later, this means a user provide bounding coordinates for the domain of interest. For applications such as basin specific benchmarking, further code development is required.

Also, further work is needed to write out newly implemented metrics to the output page.

# 3 ESMValTool

## 3.1 Overview

The ESMValTool is a community diagnostics and performance metrics tool developed to improve comprehensive and routine evaluation of Earth system models participating in the Coupled Model Intercomparison Project (Eyring et al. 2020). It consists of the core functionality (ESMValCore, which is responsible for data processing) and a diagnostic part with metrics, scientific applications and diagnostics, called recipes and translated with YAML (allowing the support of different programming languages for the diagnostics). This tool allows for evaluation of model simulations against observations, against other models or to compare different versions of the same model (Eyring et al., 2020). A working framework of ESMValTool is shown Figure 18.

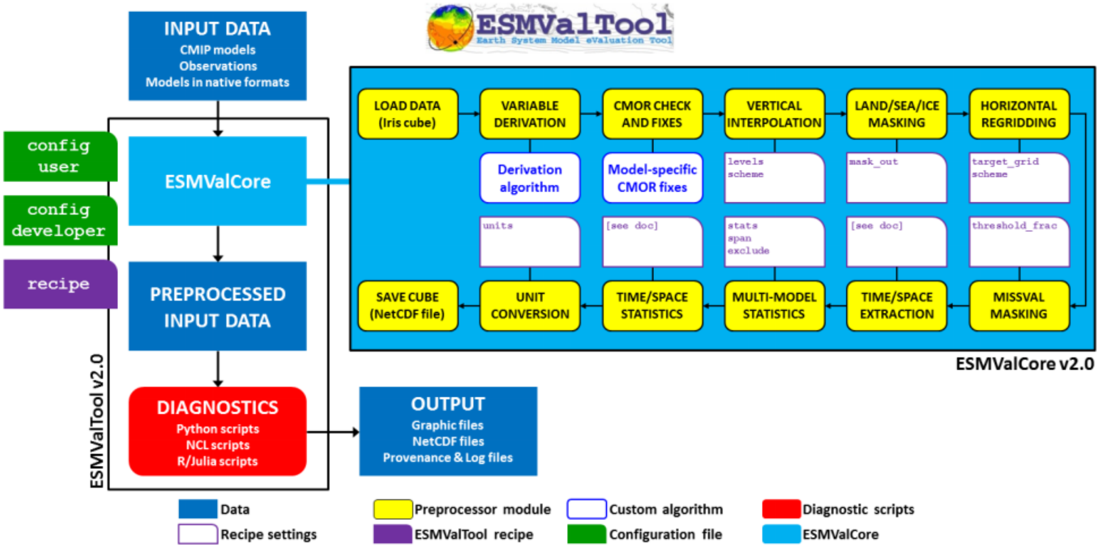


*Figure 18: ESMValTooL software framework. Source (https://docs.esmvaltool.org/en/latest/introduction.html#id3)*

## 3.2 Installation

1. Download MiniConda which is a python environment via:
   $ wget https://repo.continuum.io/miniconda/Miniconda3-latest-Linux-x86_64.sh.
2. Once downloaded, run the installation script:
   $ bash Miniconda3-latest-Linux-x86_64.sh -b -p $VSC_DATA/miniconda
3. Once the python environment is ready, run the command:
   $ conda create -n esmvaltool -c conda-forge -c esmvalgroup esmvaltool-python,
   to get the latest software from the ESMVALTOOL repository. If miniconda is already installed on the HPC skip step 1 and 2

For encountered errors during installation, check installation procedure on ESMVALTOOL webpage https://docs.esmvaltool.org/en/latest/quickstart/installation.html.

## 3.3 Pro's and Con's (caveats)

The tool cannot be used unless input data are Climate Model Output Rewriter (CMOR) compliant. CMOR outputs are metadata files which fulfils the data structure requirements of the model intercomparison projects (MIPs). In other words, input data must have specific attributes and variable names which are CMOR compliant. More information on input data structure requirements can be found on https://docs.esmvaltool.org/projects/ESMValCore/en/latest/quickstart/find_data.html?highlight=data.

This would mean that ISIMIP output needs to be restructured which was not possible within the framework of this grant.

# 4 Conclusion & Outlook

As models are become increasingly complex, there is a need for evaluation with objective comparisons against observations, for research and model improvement. In this context, this study describes the functionalities of each tool (ILAMB and ESMValTool) and caveats encountered and also document pathways for adapting these tools for model evaluation within the global water sector and potentially across the ISIMIP sectors.

Application of ISIMIP model output (mainly TWSA and river discharge) was successful using ILAMB but unsuccessful using ESMValTool due input data not being CMOR compliant. The flexibility of ILAMB in terms of code development and acceptance of external reference data made it possible to define new metric (such as KGE and NSE) in addition to the default evaluation metric and used reference data such as GRDC for benchmarking analysis. It is noted here that this newly defined metric is not visible at the output webpage directly and this can be improved in future studies. Even though ILAMB support regional benchmarking, basin specific benchmarking will require further code development since basin selection is via defining bounding coordinate rather than shapefile selection

It is envisaged here that a data CMORization tool should be developed in future studies to enable the use of ESMValTool.

This study will be beneficial for modelling teams in terms of supporting their modelling activities. Also, there is an additional advantage of model evaluation reproducibility for verification purposes. The experiences gained in this grant could form the basis of a potential model benchmark protocol for impact models through the umbrella of PROCLIAS.

# Acknowledgements

The authors thank the funding body European Cooperation in Science and Technology (COST) for the grant used for this study. We are also grateful to the ISIMIP coordination team as well as the leaders of the ISIMIP global water sector for providing technical assistance, especially Matthias Büchner to develop a script for converting GRDC data. We extend our gratitude to Wim Thiery for providing access to technical infrastructure used for this study. We would like to thank the creators of ILAMB and ESMValTool for providing the frameworks and documentation/source code. Finally, we thank the Global Runoff Data Centre (GRDC, http://grdc.bafg.de), in Koblenz, Germany, for providing the discharge data used in this study.

# References

Collier, N., Hoffman, F. M., Lawrence, D. M., Keppel-Aleks, G., Koven, C. D., Riley, W. J., Mu, M., & Randerson, J. T. (2018). The International Land Model Benchmarking (ILAMB) System: Design, Theory, and Implementation. *Journal of Advances in Modeling Earth Systems*, *10*(11), 2731–2754. https://doi.org/https://doi.org/10.1029/2018MS001354

Eyring, V., Bock, L., Lauer, A., Righi, M., Schlund, M., Andela, B., Arnone, E., Bellprat, O., Brötz, B., Caron, L.-P., Carvalhais, N., Cionni, I., Cortesi, N., Crezee, B., Davin, E. L., Davini, P., Debeire, K., de Mora, L., Deser, C., … Zimmermann, K. (2020). Earth System Model Evaluation Tool (ESMValTool) v2.0 – an extended set of large-scale diagnostics for quasi-operational and comprehensive evaluation of Earth system models in CMIP. *Geoscientific Model Development*, *13*(7), 3383–3438. https://doi.org/10.5194/gmd-13-3383-2020

GRDC. (2021). *The Global Runoff Data Centre (GRDC)*. http://www.bafg.de/GRDC

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. *Journal of Hydrology*, *377*(1–2), 80–91. https://doi.org/10.1016/J.JHYDROL.2009.08.003

Knoben, W. J. M., Freer, J. E., & Woods, R. A. (2019). Technical note: Inherent benchmark or not? Comparing Nash–Sutcliffe and Kling–Gupta efficiency scores. *Hydrology and Earth System Sciences*, *23*(10), 4323–4331. https://doi.org/10.5194/hess-23-4323-2019

Krysanova, V., Zaherpour, J., Didovets, I., Gosling, S. N., Gerten, D., Hanasaki, N., Müller Schmied, H., Pokhrel, Y., Satoh, Y., Tang, Q., & Wada, Y. (2020). How evaluation of global hydrological models can help to improve credibility of river discharge projections under climate change. *Climatic Change*, *163*(3), 1353–1377. https://doi.org/10.1007/s10584-020-02840-0

Landerer, F. W., & Swenson, S. C. (2012). Accuracy of scaled GRACE terrestrial water storage estimates. *Water Resources Research*, *48*(4). https://doi.org/https://doi.org/10.1029/2011WR011453

Müller Schmied, H., Cáceres, D., Eisner, S., Flörke, M., Herbert, C., Niemann, C., Peiris, T. A., Popat, E., Portmann, F. T., Reinecke, R., Schumacher, M., Shadkam, S., Telteu, C.-E., Trautmann, T., & Döll, P.

(2021). The global water resources and use model WaterGAP v2.2d:  model description and evaluation. *Geoscientific Model Development*, *14*(2), 1037–1079. https://doi.org/10.5194/gmd-14-1037-2021

Müller Schmied, H., Cáceres, D., Eisner, S., Flörke, M., Herbert, C., Niemann, C., Peiris, T. A., Popat, E., Portmann, F. T., Reinecke, R., Shadkam, S., Trautmann, T., & Döll, P. (2020). *The global water resources and use model WaterGAP v2.2d - Standard model output*. PANGAEA. https://doi.org/10.1594/PANGAEA.918447

Müller Schmied, H., Eisner, S., Franz, D., Wattenbach, M., Portmann, F. T., Flörke, M., & Döll, P. (2014). Sensitivity of simulated global-scale freshwater fluxes and storages to input data, hydrological model structure, human water use and calibration. *Hydrology and Earth System Sciences*, *18*(9), 3511–3538. https://doi.org/10.5194/hess-18-3511-2014

Nash, J. E., & Sutcliffe, J. v. (1970). River flow forecasting through conceptual models part I — A discussion of principles. *Journal of Hydrology*, *10*(3), 282–290. https://doi.org/10.1016/0022-1694(70)90255-6

Sutanudjaja, E. H., van Beek, R., Wanders, N., Wada, Y., Bosmans, J. H. C., Drost, N., van der Ent, R. J., de Graaf, I. E. M., Hoch, J. M., de Jong, K., Karssenberg, D., López López, P., Peßenteiner, S., Schmitz, O., Straatsma, M. W., Vannametee, E., Wisser, D., & Bierkens, M. F. P. (2018). PCR-GLOBWB 2: A 5 arcmin global hydrological and water resources model. *Geoscientific Model Development*, *11*(6), 2429–2453. https://doi.org/10.5194/gmd-11-2429-2018

Wada, Y., Wisser, D., & Bierkens, M. F. P. (2014). Global modeling of withdrawal, allocation and consumptive use of surface water and groundwater resources. *Earth System Dynamics*, *5*(1), 15–40. https://doi.org/10.5194/esd-5-15-2014

# Appendix

## A1 DATA PREPROCESSING

### Total Water Storage Anomaly

#### *Obtaining  data*

Grace twsa data in ILAMB is accessed from https://www.ilamb.org/ILAMB-Data/DATA/twsa/GRACE/ and WaterGAP twsa data is obtained from https://hs.pangaea.de/model/WaterGAP_v2-2d/watergap_22d_WFDEI-GPCC_histsoc_tws_monthly_1901_2016.nc4 .

The subsequent section was performed using cdo commands. It is noted here that for simplicity cdo piping wasn't applied here

#### *Selecting period for WaterGAP*

```
    cdo selyear,2002/2014 watergap_22d_WFDEI-GPCC_histsoc_tws_monthly_1901_2016.nc4
                                tws_2002_2014.nc
```

#### *Masking Greenland from WaterGAP*

From the GRACE twsa file, select a period with twsa distribution covering all remaining land since the grace data exclude Antarctica and Greenland. In this case April in the year 2002 as shown below

```
cdo -seltimestep,4 twsa_0.5x0.5.nc twsa_0.5*0.5_4.nc
```

Divide the file with itself to get ones in all data points

```
cdo div twsa_0.5*0.5_4.nc twsa_0.5*0.5_4.nc twsa_0.5*0.5_mask.nc
```

Latitude orientation for the twsa mask (S-N)  differ from WaterGap (N-S). The code below fixes that

```
cdo invertlat twsa_0.5*0.5_mask.nc twsa_0.5*0.5_mask_NS.nc
```

Divide WaterGap dataset by the new mask created to mask out greenland.

```
cdo div tws_2002_2014.nc twsa_0.5*0.5_mask_NS.nc tws_2002_2014_masked.nc
```

#### *Creating anomaly*

Select reference year from WaterGAP TWS dataset

```
cdo selyear,2004/2009 tws_2002_2014_masked.nc tws_2004_2009.nc
```

Take average of the reference dataset

```
cdo timmean tws_2004_2009.nc tws_2004_2009_mean.nc
```

Subtract averaged reference from WaterGAP TWS data

```
cdo sub tws_2002_2014_masked.nc tws_2004_2009_mean.nc twsa_2002_2014.nc
```

Change variable name from tws to twsa

```
cdo chname,'tws','twsa' twsa_2002_2014.nc WaterGap_twsa_2002_2014.nc
```

### Set Units

It is noted here that the time units of benchmark and model should be same to avoid errors

```
cdo settunits,days WaterGap_twsa_2002_2014.nc WG2_twsa_2002_2014.nc
```

**RIVER DISCHARGE**

***Obtaining data***

Monthly river discharge data for 12 GRDC station was requested from the Global Runoff Data Centre (GRDC). This data spans a period between 1879 to 2021. This data was then made ILAMB compliant using the code developed by Matthias Büchner (and adapted to position GRDC data on spatially on a 0.5*0.5 grid) which is hosted on GitHub (https://github.com/ISI-MIP/GRCD_convert/issues).

Daily discharge (2001-2010) from 2 models WaterGAP (with WFDEI and GSWP3 forcings) and PCR-GLOBWB (with WFDEI forcing) were downloaded from https://esgf-data.dkrz.de/search/esgf-dkrz/.

***Selecting period***

The function below was used to select the period 2001-2010 for GRDC data.

```
cdo selyear,2001/2010 grdc.nc grdc_2001_2010.nc
```

***Resampling data to monthly data.***

Daily discharge was aggregated to monthly discharge using python for all models. CDO resampling adds a time bound variable in addition to the data variable which causes ILAMB to run with errors. Exemplary code using python for WaterGAP2 with GSWP3 forcing is shown below.

```
import xarray as xr
ds=xr.open_dataset('watergap2_gswp3_nobc_hist_varsoc_co2_dis_global_daily_2001_2010.nc
')
monthly_data=ds.resample(time='1M').mean()
monthly_data.to_netcdf('watergap2_gswp3.nc')
```

***Add unit to variable***

The discharge variable (for all models) after resampling had no unit attribute and hence this was fixed using the cdo  function “setattribute” . An exemplary code is shown for the PCRGLOBWB model.

```
cdo setattribute,dis@units="m3 s-1" pcr-globwb_wfdei.nc pcr-globwb_wfdei_new.nc
```

***Calendar conversion***

PCRGLOWB model’s calendar was converted from proleptic_gregorian to 365 days due to error obtained after trail run.

```
cdo setcalendar,"365days" pcr-globwb_wfdei_new.nc pcr.nc
```

# A2 Implementation of KGE and NSE

The implementation of KGE and NSE can be found on line 710 -733 in the confrontation.py file hosted on GitHub (https://github.com/nyenah/Confrontation_Hydrological_sector) which should be replaced with the default confrontation.py file in the ILAMB core after installation

## A3 Running ILAMB

To run global benchmark analysis the code below is used.

```
   ilamb-run --config sample.cfg --model_root $ILAMB_ROOT/MODELS/ --regions global
```

To run global and regional benchmark analysis the code below is used (eg. Australia)

```
 ilamb-run --config sample.cfg --model_root $ILAMB_ROOT/MODELS/ --regions global aust
```