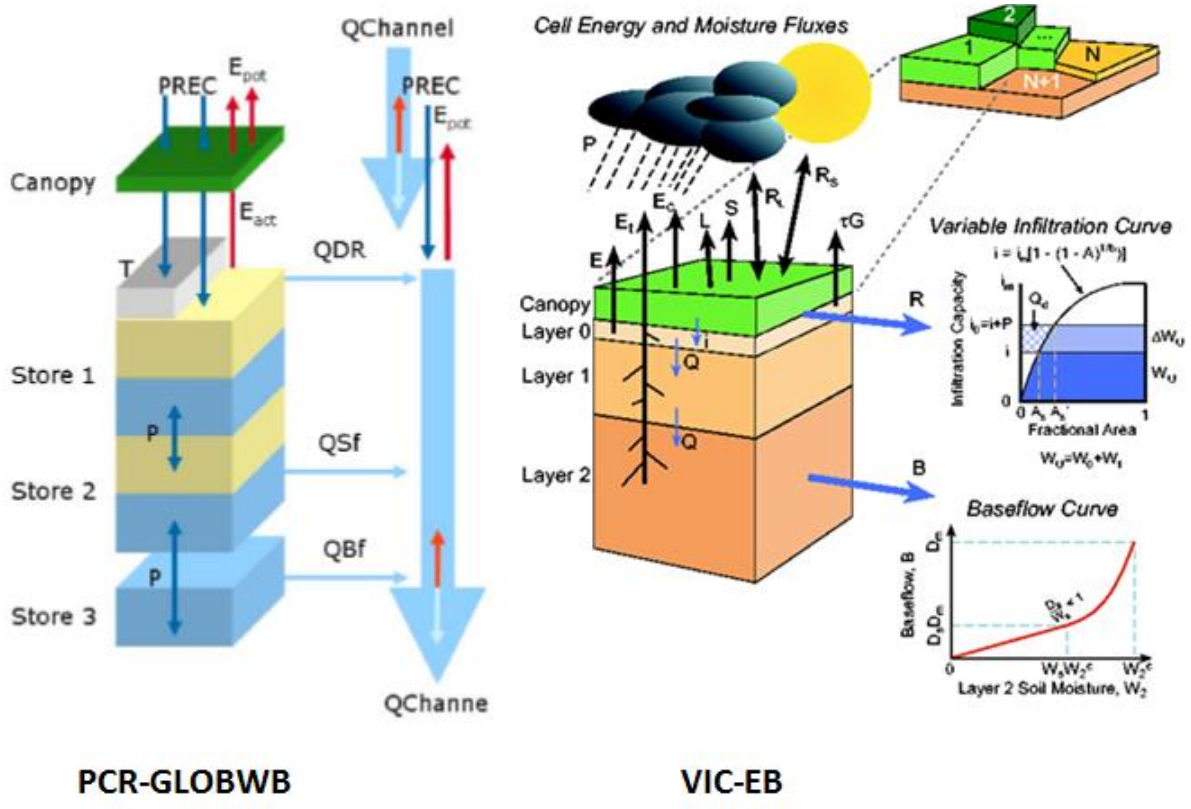
Hydrologic response modelling:

comparing the VIC and PCR-GLOBWB models



PCR-GLOBWB **VIC-EB**

Lars Killaars

Universiteit Utrecht

UNIVERSITY of WASHINGTON

Hydrologic response modelling: comparing the VIC and PCR-GLOBWB models

MSc Thesis

March - 2016

Author: Lars Killaars

Student number: 3643743

E-mail: l.killaars@students.uu.nl

First supervisor: Rens van Beek

Second supervisor: Geert Sterk

Supervisor from University of Washington: Bart Nijssen

MSc Programme: Earth Surface and Water

Faculty of Geosciences

Department of Physical Geography

Utrecht University

Abstract

Accurate information about water reserves is crucial in a world where the water demand will increase over the coming century. Global Hydrological Models (GHMs) can play a large role supporting good decision making by predicting the available water resources. GHMs have different features because of their background, each with their own strengths and weaknesses. VIC and PCR-GLOBWB are both capable of simulating the global water balance, but VIC has the potential to simulate the hydrological balance on a local scale more accurately because it solves the energy balance. This research will look at the effect that an energy balance in a GHM has on the accuracy of the output. Using the WFDEI climate forcing, PCR-GLOBWB and two versions of VIC (with and without energy balance) are used to simulate the global water balance. The model output is compared to global evapotranspiration (ET), snow water equivalent (SWE) and soil moisture datasets, as well as to discharge measurements of the Amazon, Brahmaputra, Mackenzie, Magdalena, Mississippi and the Nile. The results show that PCR-GLOBWB has higher Kling-Gupta, Nash-Sutcliffe and correlation scores for ET and equal scores to VIC for SWE. VIC, on the other hand, has higher accuracy scores for discharge in five of the six rivers and for soil moisture. The effect of an energy balance is small, as VIC-EB performs similar to VIC-WB and the results indicate that the calibration of VIC plays a larger role in the higher accuracy of VIC for the discharge and soil moisture than the energy balance. The sensitivity of PCR-GLOBWB to different climate forcing, potential ET and resolution changes was also tested. Using CRU precipitation instead of GPCC precipitation leads to significantly lower discharges, but comparable local accuracy scores. Use of the Penman-Monteith potential ET equation results in more accurate results for the entire water balance. Running PCR-GLOBWB on a higher resolution (5 arcminutes) leads to lower accuracy.

Table of Contents

List of Figures and Tables

Table

List of Appendices

List of acronyms and abbreviations

AMSR-E	Advanced Microwave Scanning Radiometer – Earth Observing System
CRU	Climatic Research Unit
E2O	EartH2Observe
ESA CCI	European Space Agency Climate Change Initiative
ET	Evapotranspiration
GHM	Global Hydrological Model
GPCC	Global Precipitation Climatology Centre
GRDC	Global Runoff Data Centre
KGE	Kling-Gupta efficiency
LSM	Land Surface Model
PCR	Same as PCR-GLOBWB
PCR-GLOBWB	PCRaster Global Water Balance
PET	Potential Evapotranspiration
NS	Same as NSE
NSE	Nash-Sutcliffe efficiency
SM	Soil Moisture
SWE	Snow Water Equivalent
VIC	Variable Infiltration Capacity
VIC-EB	VIC-Energy Balance
VIC-WB	VIC-Water Balance
WFDEI	WATCH forcing data ERA-Interim

1. Introduction

Knowledge about water resources and dynamics is important in a world where water availability is changing due to climate change. Moreover, increasing demands by demographic, land use and economic changes all contribute to driving more people towards water stress and uncertainty (Alcamo et al., 2007). In order to make accurate and good policy decisions on water related topics, it is necessary to have information about the water resources in a country. However, several countries do not have this information available.

The EU-funded EartH2Observe (E2O) program aims to produce a complete dataset of freshwater resources for the whole world. To achieve this goal, a combination of earth observation data and global hydrological models are used to determine the terrestrial water balance components (Dutra et al., 2015). This water balance can be used to reanalyse the water resources with a multi-model ensemble and to support good decision making in light of the inherent uncertainty that stems from climate forcing and model structure. As a first deliverable of E2O, the Tier1 multi-model ensemble of global hydrological models has been produced as a first estimate of the uncertainty in the water balance components (Dutra et al., 2015).

Global Hydrological Models (GHMs) use globally availably data as input to calculate the water balance on a global scale and are therefore useful to provide information to data-poor regions about their water resources. However, the outcome of these models is subject to multiple sources of uncertainty: spatial input data (such as climate data, land cover etc.), the structure of the GHM, model parameters (Müller Schmied et al., 2014) and insufficient knowledge about processes on a global scale (Beven and Cloke, 2012; Wood et al., 2011; 2012).

Multiple GHMs exist, each with a different structure, different features and different strengths and weaknesses (Sood, 2015). For this reason, an ensemble of GHMs is a strong tool to reduce the uncertainty of the results, since the strengths and weaknesses of the models may cancel each other (Gosling et al., 2011). However, the downside of an ensemble is that a certain model might be very accurate for a certain aspect, due to specific features that simulate this aspect, but these results are averaged for the ensemble results, which results in loss of accuracy for this aspect. Despite the relevance and practicality of GHMs, research about the effects of different features of specific GHMs is limited. The focus lies more on either the accuracy of the ensemble (e.g. Gudmundsson et al., 2011; Davie et al., 2013; Prudhomme et al., 2014) or the accuracy of the individual model.

Exceptions exist, for example Hurkmans et al. (2008) compared a water balance model (STREAM) to a GHM (VIC), but only briefly discussed the model structure or the actual cause of the different results of the models. Another example is the Integrated Project Water and Global Change (WATCH). An important part of this project was a large model comparison of both Land Surface Models (LSMs) and GHMs (Haddeland et al., 2011). They found some model structure related differences: models with a physically based energy balance predicted lower snow water equivalent than models with a degree day approach. This is due to snow sublimation, which is only accounted for in the energy based models. Haddeland (2011) also found that models with the Priestley-Taylor equation simulate a higher potential evapotranspiration in humid areas than models with the Penman-Monteith equation, the opposite was the case in dry areas. However, since there are also other differences in approach between the models, this cannot be solely attributed to the choice of equation.

Apart from the previously mentioned studies, research is limited on the benefits and drawbacks of specific features in GHMs. More research about the added value of certain features or which formulation of standard elements gives the most accurate results would be valuable, since it can be used to improve weak modules of a GHM and consequently improve the accuracy of the ensemble.

Since precipitation is the main source of water for the hydrological budget, this input is of critical importance for the outcome of the GHM's (Fekete et al., 2004). Several different datasets exist, each with different input sources (ground observations, satellite estimates and/or climate models) and different methods to obtain the desired resolution. Typically, the datasets agree with the main spatial and temporal patterns, but regional differences occur between the datasets (Adler et al., 2001; Fekete et al., 2004). Fekete et al. (2004) forced a water balance model with six different precipitation datasets and found that uncertainty in precipitation, results in at least the same uncertainty in the runoff. Biemans et al. (2008) evaluated the effects of different precipitation datasets for entire river basins. This study found that the average precipitation uncertainty in a basin is 30% and that this yields even larger discharge differences.

This research focuses on a comparison of two global hydrological models, PCR-GLOBWB and VIC. Furthermore, the uncertainty of the model output of PCR-GLOBWB is evaluated by looking at the effects of different model resolution, precipitation forcing and potential evaporation equations. PCR-GLOBWB was developed at the Physical Geography department of Utrecht University and is part of the ensemble used in the EartH2Observe project. VIC was developed at the University of Washington in Seattle and Princeton University and is not included in the multi-model ensemble. The main difference between both models is that VIC includes a full energy balance (Liang et al., 1994; Gao et al., 2010a), whereas PCR-GLOBWB only considers water fluxes (Van Beek and Bierkens, 2008). This research will compare results of both models, when forced with the same climatological data, by evaluating the predicted stream flows, soil moisture, snow water equivalent and evapotranspiration for different regions. The results will be used to determine the advantage of an energy balance in a GHM, as well as other strengths and weaknesses of both models that result from their model structure.

VIC can also be run in water balance only mode, VIC-WB (VIC-Water Balance). This model is, apart from the energy balance, identical to regular VIC and can be used to examine the influence of an energy balance to a model. Its results will also be used to see if there might be other differences in model structure that might lead to different results.

1.1. Research objectives

The main objective of this research is to examine the differences in model structure between VIC and PCR-GLOBWB and their effect on the calculated water balance. This will be done by answering the following research questions:

- What is the effect of the energy balance in a global hydrological model on the discharge of a major river and on local hydrological variables, such as evapotranspiration, snow water equivalent and soil moisture?
- In which climate zone will this effect be the largest: tropic, arctic, or temperate?
- How do VIC and PCR-GLOBWB perform in areas where abundant other information is available?

- Which model provides better estimates of water fluxes in regions where other information about the hydrology is not readily available?
- Will the results of VIC-WB be similar to those of PCR-GLOBWB, or are there other differences in the model structure of VIC that cause differences between the results of VIC-WB and PCR-GLOBWB?
- How sensitive is PCR-GLOBWB to changes in climate forcing, potential ET forcing and resolution?

This study will answer these research questions by comparing the results from the models to each other and to observation-based data sets on both the regional and the local scale. For the regional scale, the calculated mean discharge will be compared to measurements from the Global Runoff Data Centre (GRDC). On a more local scale, the calculated soil moisture, snow water equivalent and evapotranspiration will be evaluated and compared to global datasets. Based on the model description (see Materials and methods) and the results found by Haddeland et al. (2011), the expectation is that VIC will perform better on a local level in areas with high evaporation or snow melt, as these are energy driven processes and should be better calculated by the energy balance in VIC. For temperate regions, the energy balance becomes less important and it would be interesting to see which model performs better. VIC-WB only considers water fluxes and is therefore more similar to PCR-GLOBWB and should produce results that are more similar to PCR-GLOBWB. If this is not the case, then there are other differences between VIC and PCR-GLOBWB that might lead to different results and are worth exploring.

2. Materials and methods

2.1. PCR-GLOBWB

PCR-GLOBWB (PCRaster Global Water Balance) is a Global Hydrological Model that can be used to solve the water balance on a cell-by-cell basis and subsequently determine the distribution of water over a study area. The model is described in detail by Van Beek and Bierkens (2008), Van Beek et al. (2011) and Wada et al. (2013, 2014). The main features and features that are of interest for this study will be briefly explained in this section. A schematic concept of PCR-GLOBWB is given in Figure 1.

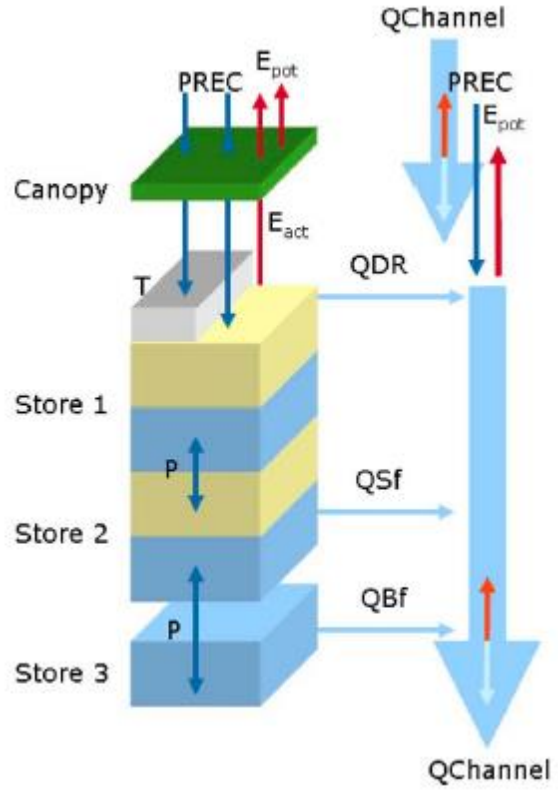


Figure 1: Schematic concept of PCR-GLOBWB. On the left are the soil compartments, two upper soil stores and a lower groundwater store. Each has a drainage component; direct runoff (QDR), interflow (QSf) and base flow (QBf). On the right side the channel is depicted, with gains (drainage and precipitation (PREC)) and losses (evaporation). (Van Beek and Bierkens, 2009)

PCR-GLOBWB is a 'leaky bucket' type of model. It calculates for each grid cell and each time step (daily) the water balance for that cell. This consists of the storage in two soil layers (Store 1 and Store 2 in Figure 1), the underlying groundwater (Store 3) and the exchange of water between the layers and between Store 1 and the atmosphere (rainfall, ET and snowmelt). Snow cover and/or a canopy with interception storage may be present, but bare soil is also possible. Snow accumulation and melt are driven by the air temperature and modelled according to the snow module of the HBV model (Bergström, 1995). This means that it is modelled on a degree-day approach, with air temperature as the driving factor for snow melt and a water holding capacity in the snow pack which delays runoff.

Not all information is specified at the same resolution as the grid size. Sub-grid variability is included in the model by several data layers with a higher resolution, such as vegetation (tall and short), open

water, soil type distribution and the fraction of saturated soil in the area based on the improved Arno scheme (Hagemann and Gates, 2003). Furthermore, the HYDRO1k Elevation Derived Database has a resolution of 1 km and is used to correct the spatio-temporal distribution of groundwater depth and to represent the snow-rain transition by introducing elevation zones and a lapse rate of 0.65°C per 100m.

Incoming water is divided into infiltration and direct runoff (QDR in Figure 1), depending on the degree of saturation of Store 1 and the water capacity of the cell. If percolation at the base of Store 1 and 2 is not possible, the excess soil moisture can drain as interflow (QSf). Drainage from the groundwater reservoir (baseflow, QBf) is calculated by a linear reservoir model (Kraaijenhof Van de Leur, 1958). The three discharge components (QBf, QSf and QDR) are routed along a river network according to the kinematic wave equations. All water surfaces are subject to open water evaporation.

PCR-GLOBWB has three routing options, depending on the accuracy needed and the computation time available. The simplest is the accutraveltime option. This option determines the velocity of a volume of water in the channel and moves it downstream according to this velocity and the length of each time step. PCR-GLOBWB also has two kinematic wave options. The simplified option uses the kinematic wave equations, and the water is limited to the channel. The last option also uses the kinematic wave equations, but the water is not limited to the channel anymore. This means that water leaves the channel and the velocity is changed due to the increased friction. This last option is the most computationally intensive, but also the most accurate.

Reservoirs are also included in PCR-GLOBWB. They are linked to the kinematic routing module. The reservoir module optimizes the release for each reservoir, based on their purpose (water supply, hydropower, flood control or 'other') and the forecast of inflow and demand along the drainage network. The outflow is updated when the actual inflow and demand deviate from the long term expectations. 94% of the total area and 95% of the total capacity of the world largest reservoirs is included; the remainder is missing due to either limited size, or lack of information (Wada et al., 2013).

Irrigation was only included in 2014, but improves the accuracy of the model in areas with high water demand (Wada et al., 2014). It includes a separation of paddy and nonpaddy crops, each with different parameters and links the irrigation with the daily soil and surface water balance. This takes the relationship into account between the irrigated water and the changes in the soil and surface water balance, which affects the soil moisture and necessary irrigation for the following days. An advantage of this method is that the soil moisture conditions and evaporation and crop transpiration are more realistically modelled. Therefore, it is not necessary to include irrigation efficiency, as this is calculated by the difference between the losses per unit crop area of evaporation and percolation and the total applied water.

An interesting feature, especially for arctic rivers, is the inclusion of a surface water energy balance into PCR-GLOBWB (Van Beek et al., 2012). This module simulates the surface temperature of freshwater bodies and this is used to predict the occurrence of river ice and those effects on the river hydraulics (wetted perimeter, Manning's coefficient etc.) and consequently the discharge.

2.2. VIC

The Variable Infiltration Capacity (VIC) model (Liang et al., 1994) was developed by the University of Washington and Princeton University and can be used as a hydrologic model and a land surface scheme. In this research, VIC was used as a global hydrological model. A comprehensive description of the model can be found in Gao et al. (2010a), but the main features and distinctive features will be described here as well.

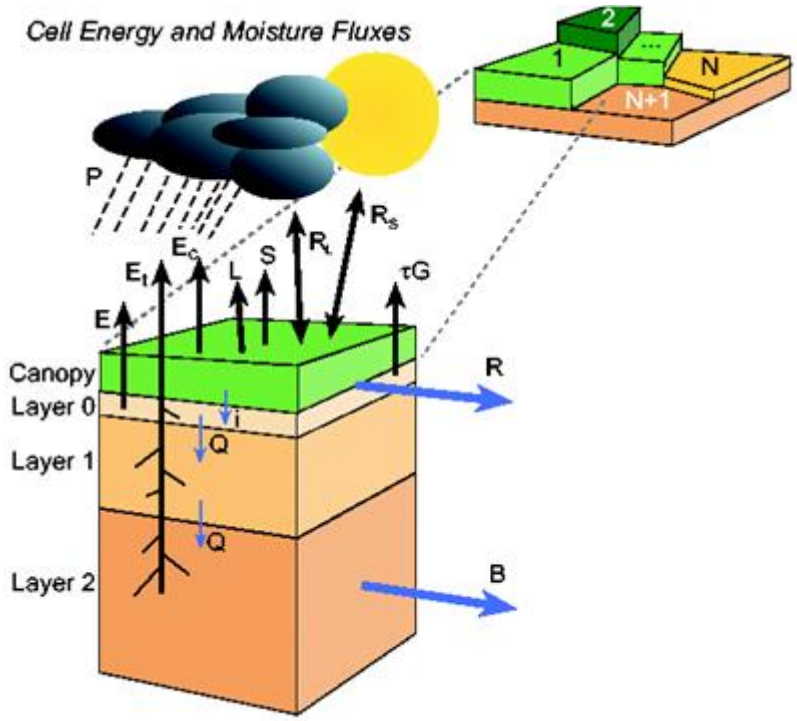


Figure 2: Schematic of the VIC 3L model with mosaic representation of vegetation coverage. Included are the different soil storage layers and moisture fluxes that are calculated or used by VIC. P is precipitation, Ec is evaporation from the canopy layer, Et is evapotranspiration from the vegetation, E is evaporation from the soil layer, L is latent heat flux, S the sensible heat flux, Rl is the incoming and outgoing long wave radiation, Rs is the incoming and outgoing short wave radiation and tG is the ground heat flux. The discharge components are indicated by R (runoff), B (baseflow), Q (percolation) and I (infiltration). Source: Gao et al., 2010a.

VIC can also be classified as a 'leaky bucket' model (Figure 2), as it also balances the water budget in a cell and creates runoff from the excess water. Furthermore, the full VIC model also balances the surface energy budget by solving for the surface temperature that balances the surface energy fluxes (sensible heat (S in Figure 3), ground heat (tG), ground heat storage and latent heat (L)) with the net shortwave(Rs) and longwave (Rl) radiation. By obtaining the surface temperature, surface processes that are energy related, for example snow melt or evaporation, are more accurately modelled. However, the downside is that this requires a sub-daily time-step, with the appropriate meteorological input data, and results in a longer computation time. The version of VIC with the energy balance is called VIC-EB in the rest of this study. VIC-WB only considers the water balance and assumes that the surface temperature is equal to the air temperature when the grid cell is snow-free. In the case of snow, the snow model operates at a sub-daily time step and solves the full energy equation at the snow surface. Overall, this still results in a significant saving of computation time, but at the cost of accuracy.

The energy balance is also included in the snow module of VIC. This module is schematically explained in Figure 3. Since the scale of snow processes is smaller than the resolution of VIC in this research, sub-grid variability is taken into account. Grid cells are divided in multiple elevation bands, which have a number of land use classes. Therefore, each grid cell has a certain number of tiles with characteristic properties. Each tile runs the snow module separately and the results are recorded as the area-averaged values of all the tiles within a grid cell.

Snow is simulated in VIC using a two layer snow pack, with energy exchange only occurring within the surface layer. The snow is also subject to compaction and ageing, resulting in a higher density and lower albedo respectively. Snow can also be intercepted by the vegetation, according to a efficiency ratio between precipitation and precipitation interception (taken as 0.6, Storck et al., 2002). The maximum storage is based on the LAI of the vegetation cover. Meltwater can be stored, both in the canopy and in the snowpack on the ground, and will only contribute to runoff once the water holding capacity has been exceeded.

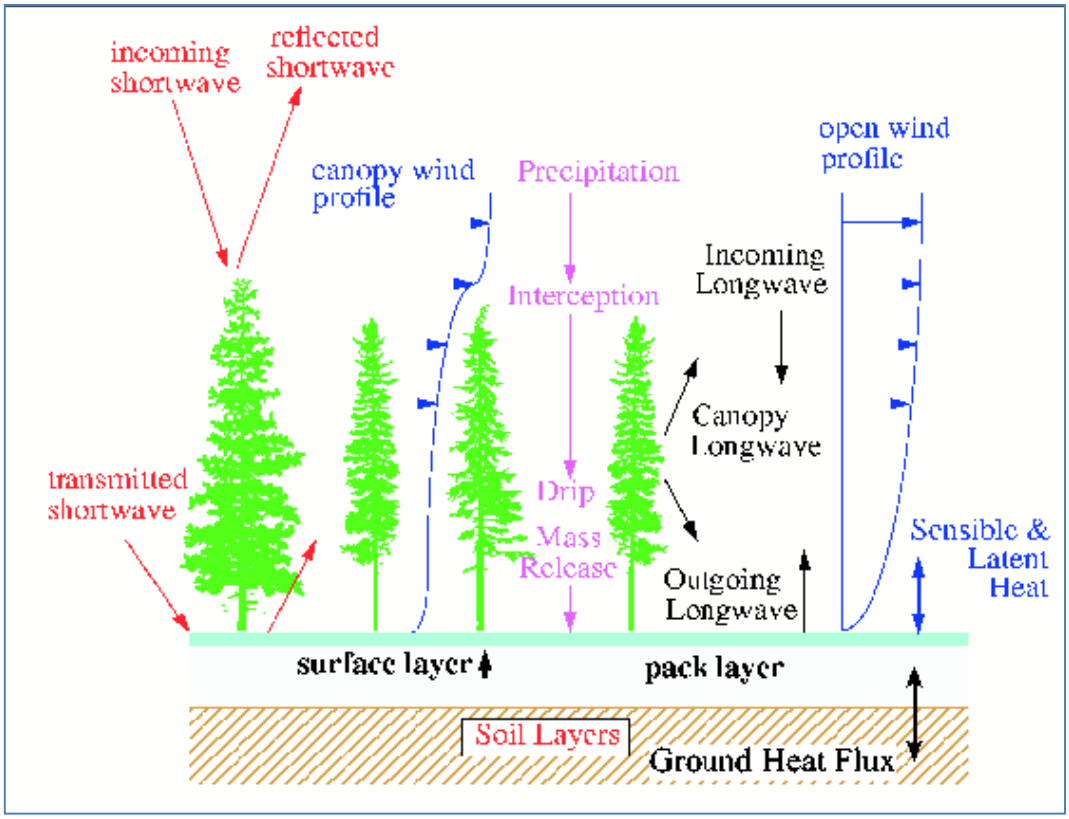


Figure 3: Schematic of snow accumulation and ablation processes in the VIC snow model. Source: Gao et al., 2010a.

As shown in Figure 2, each grid cell can have multiple tiles, each with its own vegetation characteristics. The soil characteristics are the same for each tile within a grid cell. Similar to the snow module, soil moisture, infiltration, water movement between the soil layers and runoff are calculated separately for each tile and then fractionally summed for the whole area. This is also done for the energy fluxes. Lakes and wetlands are modelled in the same way. They are assigned a tile with appropriate lake or wetland characteristics and that tile is added to the vegetation mosaic (Bowling and Lettenmaier, 2010). This option can be turned on or off, as it requires additional information to function. This study does not use this function.

The routing module of VIC (RVIC) is separate from the water budget calculations in each cell. VIC calculates the runoff for each cell and a separate routing model (Lohmann, et al., 1996, 1998) is used to transport the runoff and baseflow to the channel and through the channel network. Once the water reaches the channel, it is no longer part of the water budget and cannot re-enter the local water balance. The routing model uses a time-invariant velocity and diffusivity to calculate a unit hydrograph between each upstream grid cell and one or more downstream locations of interest.

A generic reservoir module was implemented in the routing model by Haddeland et al. (2006a). Similar to PCR-GLOBWB, the model forecasts the reservoir inflows for the next year. The optimum reservoir release is determined based on these forecasts.

VIC also has an irrigation component. The main function of this component is to avoid vegetation stress due to limited water availability (Haddeland et al., 2006b). Irrigation begins when the available soil moisture is below the level where transpiration is possible and continues until the soil reaches field capacity. Water needed for irrigation can be extracted from the local river runoff, or from a reservoir or another source in the basin. In this mode, the water availability defines the total amount of irrigation possible. The irrigation module can also be used to calculate the total water demands. In this mode the water is freely available and the irrigation can always continue to field capacity.

2.3. Important differences between VIC and PCR-GLOBWB

The aim of VIC is similar to that of PCR-GLOBWB, to calculate the water budget in a grid cell. However, by looking at the schematic figures of PCR-GLOBWB (Figure 1) and VIC (Figure 2), differences between the structures of both models become visible immediately. The first distinction is the amount of surface energy fluxes calculated by VIC compared to PCR-GLOBWB. Together, these fluxes (S, L, R_s, R_L and τG in Figure 2) make up the energy balance of VIC as described above. The direct result of this energy balance is that the surface temperature is more realistically modeled than it is with the assumption that the surface temperature is equal to the air temperature. Especially during clear conditions the difference between the surface and air temperature is large as the surface heats and cools faster than the air above the surface. This should be visible in the output of the models in the fact that snow melt and evaporation processes should be more accurately modelled on the local scale. However, the energy balance also adds more complexity. The physics are more directly represented with an energy balance, but the model also needs more detailed input data. If there is an uncertainty in the extra input data, the same would be true for the results.

Apart from the energy budget, the snow module of VIC is more advanced than the module of PCR-GLOBWB. PCR-GLOBWB uses the snow module of the HBV model (Bergström, 1995), a relatively simple degree day melt model that only deals with snow melt. VIC on the other hand also includes processes such as compaction, ageing and snow interception. Moreover, VIC also uses the energy balance for the melt and accumulation processes.

Evaporation is also calculated differently in both models. PCR-GLOBWB uses the Hamon equation coupled with crop factors for different vegetation types to calculate the reference potential evapotranspiration (PET). Bare soil evaporation can either be forced by the meteorological data (ERA40) or calculated by converting the reference evapotranspiration to reference soil evaporation and transpiration. It is also possible to force the model with the reference PET. In this case any pre-existing dataset of PET can be used. This study used the Hamon equation to calculate the reference

PET. In order to evaluate the differences between both methods, one simulation of the Amazon basin used the reference PET calculated by the Penman-Monteith equation.

VIC considers three types of evaporation: evaporation from the canopy layer, transpiration from the vegetation and bare soil evaporation. Penman-Monteith is used to calculate the potential evaporation, which is transformed to actual canopy evaporation and vegetation transpiration using vegetation-specific resistance coefficients. Bare soil evaporation is similar for saturated soils in both models, with the hydraulic conductivity as the maximum limit, otherwise it is the calculated evaporation rate. The Arno formulation is used for unsaturated soil. VIC uses the Arno formulation of Franchini and Pacciani (1991) while PCR-GLOBWB uses the revised Arno scheme of Hagemann and Gates (2003).The main change in the revised Arno scheme with respect to evaporation is the introduction of the minimum soil water capacity. In the old scheme, the minimum soil water capacity was set at zero, while in the improved scheme this can be chosen according to the type of soil (Hagemann and Gates, 2000). The fractional saturation curves are also improved. These curves are used to determine the distribution of soil water capacities in a gridcell and therefore influence the infiltration and runoff within a gridcell.

Another difference in model structure between both models is the amount of sub grid variability each model takes into account, essentially improving the resolution of the model. VIC offers more versatility in the amount of vegetation/land use tiles that can be chosen for each grid cell. The user can specify a 'N+1' number of tiles, meaning that the amount of vegetation tiles is only limited by the amount of information present for each class and the effort that the user is willing to put into the parameterization phase. PCR-GLOBWB has six vegetation classes: natural vegetation, rain-fed crops and irrigated crops, all three subdivided into tall and short vegetation. The separation is made based on the rooting depth of the vegetation. Tall vegetation reaches the second soil layer and short vegetation only draws water out of the first soil layer. In VIC, the rooting depth can also be specified for each vegetation tile.

Even though the vegetation can have many different characteristics in each grid cell in VIC, the soil parameters are kept constant within each gridcell. PCR-GLOBWB, in contrast, offers sub grid variability for the soil type along with the characteristics for each type. Moreover, due to the improved Arno scheme in PCR-GLOBWB, the variability of sub grid soil saturation is better accounted for (Hagemann and Gates, 2003).

VIC does not have a module similar to the river temperature module from PCR-GLOBWB. As a result, PCR-GLOBWB might produce more accurate results for open water evaporation due to the more accurate river temperature. Moreover, in arctic regions PCR-GLOBWB has the possibility to model river ice with this module. The occurrence of river ice and the decrease of discharge in rivers with ice jams in PCR-GLOBWB might be one of the causes for different results in the arctic case study.

Another difference is the amount of discharge fluxes. PCR-GLOBWB calculates three different flows to the channel: direct runoff, interflow and baseflow. VIC only differentiates direct runoff from baseflow. Because the water balance remains equal, i.e. there is no water lost or created, it will not have significant consequences for the discharge. It might reduce the peak of discharge in VIC, as the water will have to travel to the lowest soil layer before it can become baseflow. In PCR-GLOBWB, the middle soil layer can also drain to the river, which could lead to a more concentrated peak in discharge.

2.4. Study areas

The study uses seven domains, consisting of a global domain with a spatial resolution of 0.5-degree (30 arcminutes, approximately 50km, depending on the latitude) as well as six specific regional case study areas. In PCR-GLOBWB the regional case study areas were simulated on a 0.083 degree resolution (5 arcminutes, approximately 10km) to compare them to the low resolution results of the global comparison. Three of these regional case study areas are EartH2Observe case study areas: the Magdalena-Cauca catchment in Colombia, the Nile catchment in Eastern Africa and the Brahmaputra catchment in Bangladesh/India. The other three, which represent the different climate regions of the world, were selected because they are large, and have sufficient information available for model evaluation: the Amazon in South-America (tropic), the Mississippi in the US (temperate) and the Mackenzie in Canada (arctic). The case study areas are described below, each with their unique characteristics that could lead to differentiation between the model results of VIC and PCR-GLOBWB.

Figure 4: Main river basins in the world. Case study regions are marked by a red circle. Modified from United Nations Environment Program (UNEP); World Conservation Monitoring Centre (WCMC); World Resources Institute (WRI); American Association for the Advancement of Science (AAAS); Atlas of Population and Environment, 2001.

For areas with large amounts of available information, GHMs might seem less useful at first glance. However, these areas can be used to evaluate the models on a regional scale and possibly calibrate them. If models perform well in these regions, it may be inferred that they perform similarly in other areas, as long as the quality of the input is the same. Hence, if the models are forced with globally available data and perform adequately for the three information-rich regions, it may be expected that the models will perform adequately in the EartH2Observe case study areas as well.

Table 1: Statistics for the six different rivers used in this research. Source length and height difference: Penn, 2002

River Basin	Climate	Height difference (m)	Length (m)	Average discharge	EartH2Observe study area	Discharge station location	References discharge
Amazon	Tropical	5,170	6,276	203,000	No	Jatuares	Martinez et al., 2009
Brahmaputra	Monsoon	5,210	2,897	20,000	Yes	Bahadurabad	Gain et al., 2011
Mackenzie	Arctic	156	1,802	9,910	No	Arctic Red River	Emerton et al., 2007
Magdalena	Tropical	3,685	1,600	7,200	Yes	Calamar	Gottschalk et al., 2015
Mississippi	Temperate	450	3,734	16,792	No	Vicksburg	Goolsby, 2000
Nile	Tropical and arid	2,700	6,853	2,830	Partly (Blue Nile catchment)	Dongola	Nile Basin Initiative: www.nilebasin.org

Each of the rivers has specific features that could provide different results between VIC and PCR-GLOBWB. The Magdalena and the Amazon are both in a tropical climate, resulting in high evaporation and precipitation rates during the year. Moreover, the Amazon has an immense drainage basin with a very high average discharge. The size as well as the climate should provide a challenge for both PCR-GLOBWB and VIC to obtain accurate model results. The Brahmaputra is the largest glacier fed river system of Bangladesh. The river basin (Figure 4, area 19) lies in an area with intensive water use, either for agriculture or for economic and demographic demands. Discharges are especially high during the spring, when the Himalayan snow melts and during the summer, which is the monsoon season in Bangladesh. The high water use, together with the monsoon, high temperatures and snow melt rates, makes this an interesting case study to test the evaporation, snow melt and water use components of the models.

The Mackenzie is an arctic river where snowmelt is very important for the discharge. A distinct feature of the Mackenzie that is hard to model accurately is river ice. River ice is a key factor in the occurrence of extreme hydrologic events in arctic regions (Goulding et al., 2009). It causes ice jams, which force the river to rise behind it. This can result in flooding but also in high discharge events when the jam clears due to high temperature or enough water pressure behind it (Beltaos, 2008). The large fraction of snow melt in the discharge of the Mackenzie could also be a distinction between both models for this area. The Mississippi is the main river of the American land and flows through a highly productive farming region, where about 58% of the total catchment is cropland (Goolsby, 2000). This means that it has a high water use component and this could cause different results for the two models. The last river is the Nile. It has a diverse catchment, with tropical rainforest in the south and the deserts of Sudan and Egypt in the north. This means that along the whole river evaporation rates are expected to be high, but only the southern part of the basin receives a significant amount of precipitation. Moreover, a large amount of water is used for irrigation, further reducing the discharge of the river. These two characteristics of the basin could provide a challenge for VIC and PCR-GLOBWB.

2.5. Data

Both VIC and PCR-GLOBWB are constructed to solve the water balance (Eq. 1) for each cell. In Eq. 1, ΔS is the change of storage in a cell (either groundwater, soil moisture or snow water equivalent (SWE)) over a certain time Δt. The terms on the right-hand-side are accumulated fluxes over the interval Δt, where P is the precipitation, E is the evapotranspiration (ET) and Q is the runoff that the

cell generates. To evaluate VIC and PCR-GLOBWB, this research will look at all components of the water balance.

$$\frac{\Delta S}{\Delta t} = P - E - Q \qquad \text{Eq. 1}$$

The precipitation is part of the climatic forcing of both models. This research will use data from the WATCH-Forcing-Data-ERA-Interim (WFDEI) meteorological forcing dataset (version 3, Weedon et al., 2014), as the data from this project is also used in the EartH2Observe ensemble. There are two precipitation datasets available in the WFDEI: the CRU (Harris et al., 2014) and the GPCC (Schneider et al., 2014). Both will be used in this research, but for the VIC-PCR-GLOBWB comparison the GPCC will be used, as this dataset includes more observation points. For the PCR-GLOBWB comparisons, a global simulation is made of PCR-GLOBWB on 30 arcminutes resolution with the CRU precipitation, to allow for comparison with the EartH2Observe project. The models will be forced with data from the whole period (1979-2010) and the results from the final years (2005-2010) will be used to compare the models. Only the final period will be evaluated, to allow for a spin-up of the models and because the quality of the comparison datasets is higher during this period than during the period before.

Table 2: Details of the global input datasets used in this research.

Name dataset	Parameter	Spatial resolution	Temporal resolution	Unit	Reference
CRU – TS	Precipitation	0.50 degrees	Daily	m/day	Harris et al., 2014
GPCC	Precipitation	0.50 degrees	Daily	mm/day	Schneider et al., 2014
FLUXNET-MTE	ET	0.50 degrees	Monthly	mm/day	Jung et al., 2009
ERA-Interim	ET	0.70 degrees	Daily	m/day	Dee et al., 2011
GRDC	Discharge	1 station per river	Monthly	m^3/s	http://www.bafg.de/GRDC
ASMR-E	SWE	0.25 degrees	Monthly	mm	Tedesco et al., 2004
ESA CCI	Soil moisture	0.25 degrees	Monthly	m^3/m^3	Liu et al., 2011; 2012; Wagner et al., 2012

The ability to simulate the evapotranspiration of the models will be tested by comparing the results with upscaled measurement data. This research uses the evapotranspiration estimates from the FLUXNET-MTE data set (Jung et al., 2009). This data set provides a gridded, global estimate of fluxes of sensible and latent heat from the surface to the atmosphere. This can be converted to ET, since the latent heat is the energy version of the ET flux. It is based on the global dataset of eddy covariance measurements, the FLUXNET database (Oak Ridge National Laboratory Active Archive center, 2015). Jung et al. (2009) developed a new model for the upscaling of the measurements to a global grid. This model was based on the measurements of 178 tower sites, mostly located in North America and Europe. This means that the accuracy of the dataset is higher in this part of the world and a large amount of gridcells are dependent on the interpolation between the sites. Jung et al. (2009) also provides a standard deviation map of this dataset, which also indicates that the possible error is largest for the tropics. Because large negative correlations were found when comparing both models to the Fluxnet dataset in parts of the Amazon and tropical Africa (amongst other regions), an extra ET dataset was used to gain more insight in these regions. The ERA-Interim ET dataset was

chosen for this purpose (Dee et al., 2011). ERA-Interim is the latest global atmospheric reanalysis produced by the European Centre for Medium-Range Weather Forecasts. ERA uses conventional observations coupled with satellite data and is therefore less dependent on limited observation points (Dee et al., 2011).Mueller et al. (2011) performed a cluster analysis with several ET datasets and found that Fluxnet and ERA-Interim do not have a large degree of association between them. This means that these datasets show specific spatial patterns that are less correlated to each other than to several other ET datasets. This could lead to different results when the models are compared to the Fluxnet ET or to the ERA-Interim ERT. Details about the resolutions and units of these datasets can be found in Table 2.

The monthly averaged discharge is compared to measurements obtained from the Global Runoff Data Centre (GRDC) (available from http://www.bafg.de/GRDC). All rivers had at least several years of monthly averaged discharge measurements available in the GRDC, but not all reached 2010. The Bangladesh only had discharge data available in the GRDC from 1985-1992. Data from the Bangladesh Water Development Board was used for this river, as this consists of a longer period (1985-2010)(obtained from ffwc.gov.bd on 01/12/2015). These are the observation locations for each river and the length of the timeseries used:

- Amazon at Jatuares: 1992-2010. There was a longer dataset available for the Amazon, but the models needed a longer spinup.
- Brahmaputra at Bahadurabad: 1985-2010
- Mackenzie at Arctic Red River: 1982-2010
- Magdalena at Calamar: 1980-1990
- Mississippi at Vicksburg: 1980-2010
- Nile at Dongola: 1982-2002

The difference in runoff generation was also evaluated. For each cell the evapotranspiration was subtracted from the precipitation. This results in the total amount of water available for runoff generation for both models. This does not directly equal the amount of runoff in each cell, but given the assumption that over a long period the change in storage is negligible to the amount of runoff simulated in that same period, it gives an indication of the runoff generated by that cell. Since there are no global datasets that measure this parameter, the obtained results will only be compared to each other.

The storage part of the water balance equation will be evaluated using two parameters: the soil moisture in the top layer and the snow water equivalent (SWE). The results of both parameters will be compared to remote sensing data, the ESA CCI Soil Moisture dataset and the AMSR-E monthly snow water equivalent products. The soil moisture product is a combination of four passive soil moisture retrievals (SMMR, SSM/I, TMI and AMSR-E) and two active (ERS AMI and ASCAT) (Wagner et al., 2012, Dorigo et al., 2015). This combination leads to an accuracy similar to the best performing input product, but with an increased temporal resolution and is therefore preferable to a singular product (Liu et al., 2011, Liu et al., 2012). This product has three main drawbacks. First, it is only checked with other observations for the northern hemisphere (Dorigo et al., 2015), second, it only looks at the top part of the soil and third, it does not give any results when there is snow or permafrost in the area. This study compares the soil moisture of the first soil layer of PCR-GLOBWB and VIC with the satellite data. For PCR-GLOBWB this layer has a depth of 13-30cm and for VIC this

depth is 10cm. Satellites only provide information about the top part of the soil (depending on the soil up to 10cm deep), meaning that not the exact same thing is compared. Especially after weather events there might be a difference between the models and the satellite data. However, because the soil moisture data is monthly averaged before the comparison, this effect is expected to be negligible.

The AMSR-E instrument does not only provide data for soil moisture, its signal can also be used to estimate the SWE (Tedesco et al., 2004). There are several of these so called snow products, for this research the monthly snow water equivalent data set is used as both models can simulate the SWE of each grid cell. Details about the resolutions and units of these datasets can be found in Table 2.

Table 2 shows that not all input and comparison data has the same spatial or temporal resolution as the resolution of the models for the different comparisons. This means that in order to use these datasets the resolution has to be changed. All comparison data was first scaled to the correct resolution (either 30 or 5 arcminutes) and then compared with the results of VIC and PCR-GLOBWB.

Other input for both models are the soil characteristics. The FAO Digital Soil Map of the World (FAO, 2003) was used as input for both models. This map gives the dominant soil for each grid cell and based on the soil type the parameters used in the model are determined. VIC used the soil parameter file from Zhou et al. (2016), which was calibrated to ensure realistic reservoir behaviour. They disaggregated the streamflow to produce spatial runoff fields, which were then used to calibrate the model parameters of VIC. Consequently, five of the six basins were calibrated (the Magdalena is the only uncalibrated basin). All basins in PCR-GLOBWB are uncalibrated.

2.6. Model configuration

PCR-GLOBWB was configured the same for all simulations. The irrigation, water demand and reservoir functions were all activated. Both VIC models and PCR-GLOBWB were run two times for the model structure comparison. The first run was used as a spin-up to obtain accurate base flow rates and water storage (soil, groundwater and snow water equivalent). The second run used the final results of the models as initial conditions. PCR-GLOBWB was run during the first run with the accuTravelTime routing option and the second time with the simplifiedKinematicWave.

Only PCR-GLOBWB was used for the other runs. Since the focus of comparison was more on the last period of each simulation (2005-2010), only one run was done. For all runs, except the Evapotranspiration comparison, the runtime was split in two, to make a distinction in the routing. The routing from 1979 to 2004 was calculated with the accuTravelTime function, instead of the kinematic wave. This period was used as a spin-up. To obtain the most accurate routing results possible for the period used for the comparison, the second period (2005-2010) was simulated with the kinematic wave equations. This was done because the kinematic wave function needs more calculation time than the accuTravelTime function. Some EartH2Observe catchments only had discharge data available for the period before 2005. To obtain the most accurate results for these catchments as well, PCR-GLOBWB was run using the simplified kinematic wave (Brahmaputra and the Nile) or the kinematic wave (Magdalena-Cauca) function.

In order to obtain the most accurate comparison results, VIC was configured as similar as possible to PCR-GLOWB. This study used the version 4.2.a with irrigation activated (available at https://github.com/UW-Hydro/VIC/tree/support/VIC.4.2.a.irrigation on 04/03/2016). VIC was run in

water balance (VIC-WB) and energy balance (VIC-EB) mode. Apart from the energy balance, no other differences in configuration between both VIC models occurred. VIC-WB was run twice for 1979 to 2010, with the final conditions of the first run as the initial conditions of the second run. VIC-EB was run only once, but used the same initial conditions as the second VIC-WB simulations. To obtain the discharge at the measurement stations for the rivers, the streamflow routing model RVIC was used.

2.7. Comparison scheme

In order to answer the research questions, several separate comparisons were made. The main focus is on the comparison between both VIC versions and PCR-GLOBWB to evaluate the differences in model structure between these models. This is the only comparison included in the main part of the thesis. The other three comparisons are discussed in Appendix 1. Three different comparisons were performed to test the sensitivity of PCR-GLOBWB to different climate input data, resolution of the model and reference evapotranspiration method. The different properties of each comparison are shown in Table 3. All input and comparison data was first converted to the used resolution (either 5 or 30 arcminutes) before it was used or compared with the results of PCR-GLOBWB

Table 3: Different comparison schemes and their properties. For the potential evapotranspiration equations PCR-GLOBWB simulated the Amazon basin using the Hamon equation and the Penman-Monteith equation. Only the model structure comparison is discussed in the results and discussion. See Appendix 1 for the results of the other three comparisons. The six case study regions are the basins of the Amazon, Brahmaputra, Mackenzie, Magdalena, Mississippi and the Nile.

Comparison	Model	Resolution	Climate forcing	Region
Model structure	VIC-EB, VIC-WB and PCR-GLOBWB	30 arcminutes	GPCC	Worldwide
Climate forcing	PCR-GLOBWB	30 arcminutes	GPCC and CRU	Worldwide
Resolution	PCR-GLOBWB	5 and 30 arcminutes	CRU	6 case study regions
Potential Evapotranspiration equations	PCR-GLOBWB	5 arcminutes	CRU	Amazon basin

2.8. Nash-Sutcliffe and Kling-Gupta efficiency scores

The comparison of the discharge, evapotranspiration, soil moisture and snow water equivalent was done by comparing the coefficient of determination, Nash-Sutcliffe efficiency and Kling-Gupta efficiency. The Nash-Sutcliffe efficiency (NSE) (Nash and Sutcliffe, 1970) is one of the most widely used criteria for the comparison of hydrological models. It can be interpreted as a classic skill score (Murphy, 1988), meaning that it compares the 'prediction skill' of a model to a certain baseline. In the case of NSE, this baseline is the observations. It uses the following equation:

$$\mathrm{NSE} = 1 - \frac{\sum_{t=1}^{n}(x_{s,t} - x_{o,t})^2}{\sum_{t=1}^{n}(x_{o,t} - \mu_o)^2} \qquad \text{Eq. 2}$$

Where n is the number of timesteps, $x_{s,t}$ is the simulated value at timestep t, $x_{o,t}$ is the observed value at timestep t and μ_o is the mean of the observed values.

The NSE value is dimensionless and ranges from -infinity to 1, where 1 means that the model predictions match the observed values perfectly. An NSE value of 0 means that the model predictions perform no better than when the mean of observations is used as the predictor. A value

less than 0 means that the mean of the observations is a better predictor than the model. Even though the NSE is an easy-to-use predictor for model performance, it has its drawbacks. Modifications have been proposed by McCuen and Snyder (1975), Krause et al. (2005), Gupta and Kling (2009) and other authors. Especially the fact that NSE uses the observed mean as a baseline is criticized, because this means that extreme values and seasonality have a disproportionally large influence on the eventual score of the NSE.

This research also uses an alternative measure, the Kling-Gupta efficiency, to complement the evaluation based on NSE values. Gupta and Kling (2009) created this alternative model performance criterion by decomposing the NSE into three components: the correlation, the variability error and the bias error (see eq. 3). They found that when you maximize the NSE value of model predictions, the variability will be underestimated. To improve this, they introduced an alternative model performance criterion, the Kling-Gupta efficiency (KGE). KGE consists of the same three components, but uses the Euclidian distance of the components to calculate the efficiency. As a consequence, the model calibration process is improved, as each component has to be calculated before the KGE can be calculated, resulting in a clear view of the origin of the model error.

$$KGE = 1 - \sqrt{(r-1)^2 + (\alpha-1)^2 + (\beta-1)^2}$$

Where:

r = linear correlation coefficient

$$\alpha = \frac{\sigma_s}{\sigma_o}$$

$$\beta = \frac{\mu_s}{\mu_o}$$

Eq. 3

This study uses the NSE, KGE and r scores to determine the accuracy of the models compared to the comparison data. The discharge results consist of time series for a certain location and from this the NSE, KGE and r^2 scores can be computed easily. The soil moisture, snow water equivalent and the evapotranspiration are spatially plotted for the entire area. Each cell has.a value for every timestep (a month) and the NSE, KGE and correlation scores are calculated for each cell. This means that all the predicted values for each cell are compared to the monthly values for the measurements. This results in a large grid with NSE, KGE and correlation values, one for each cell for each variable. For comparison, these scores were plotted in spatial plots and in boxplots.

The model results were not only compared to the comparison datasets, but also to each other. The NSE, KGE and correlation scores were also calculated in a similar manner as above, but using PCR-GLOBWB as if it were the observations. This produces spatial plots which show the areas where PCR-GLOBWB and VIC differ the most.

3. Results of the VIC-PCR-GLOBWB comparison

Here, the results of both VIC models and PCR-GLOBWB are compared to each other. As mentioned in the comparison scheme (see Table 3), both models were forced with the GPCC precipitation data and the simulation was done on a resolution of 30 arcminutes.

3.1. Discharge

Figure 5 shows, for each different model/setup, the average monthly discharge for each river. Keep in mind that VIC was calibrated for all studied rivers, except the Magdalena. The location of the discharge measurements and the corresponding KGE, NSE and r^2 scores of the monthly discharge timeseries can be found in Table 4. Figure 6 shows the same scores, but in a graph to allow for easy comparison between the different models and rivers. To clarify, the scores in table are calculated based on the monthly discharge timeseries observed by the GRDC and simulated by PCR-GLOBWB and VIC and not on the average monthly discharge shown in Figure 5.

It can be noted that each river is simulated accurately by at least one model, but there is not one model that performs well on every river. The river that is simulated most accurately by the most models is the Brahmaputra. All three models (PCR-GLOBWB, VIC-EB and VIC-WB) have the highest NS and high KGE scores for this basin. Especially VIC performs very well, with NS scores of 0.92 (VIC-EB) and 0.90 (VIC-WB). However, as the average discharge shows, the simulations are still not perfect. VIC-EB simulates the peak months during the summer well, but overestimates the dryer months. VIC-WB underestimates the peak months, but performs the best during the dry months. PCR-GLOBWB underestimates the peak during the summer by around 20,000m^3/s and overestimates the dry season. This leads to lower scores than the VIC models with a NS of 0.74 and a KGE of 0.56. The r^2 of all models are high; both VIC models score 0.94 and PCR-GLOBWB 0.88.

Another river that is simulated well in general is the Mackenzie. Again, both VIC models have the highest scores for this river (Table 4). Although the melting season in VIC starts earlier than observed by the GRDC and simulated by PCR-GLOBWB, the peak discharge in VIC corresponds to the observed values better than the PCR-GLOBWB results do. All models have trouble mimicking the behavior of the Mackenzie during the rest of the year. Both VIC models drop too fast after the melt peak and underestimate the tail during the summer. This gap closes later in the season and during the winter there is only a small underestimation. PCR-GLOBWB's drop is slower than the observed values and therefore ends up overestimating the discharge at the end of the year. The NS, KGE and r^2 scores confirm these findings as both VIC models score the highest, with VIC-EB slightly outperforming VIC-WB, followed by PCR-GLOBWB

For the other rivers there is only one model (either VIC-EB, VIC-WB or PCR-GLOBWB) that performs well, whereas the others score lower. An example is the Magdalena; both VIC models score negative for the NS and have very low r^2 and KGE scores for this basin, PCR-GLOBWB scores better. VIC has trouble with the trend of the Magdalena, which results in the low r^2 scores and thus also in low NS and KGE scores. PCR-GLOBWB performs better and due to the good r^2, the KGE score remains relatively high.

Both VIC models perform better for the Amazon than PCR-GLOBWB. The latter has trouble with the amount of water present in the basin, as its peak discharge is almost 70,000 m^3/s lower than the observed value. VIC also does not reach the observed peak discharge, but the gap is smaller.

However, the timing of the peak is a month early in VIC, resulting in an underestimation for the rest of the year until the 'dry' season. PCR-GLOBWB has a correct timing of the peak discharge, but there is limited seasonality in the calculated discharge. This means that the gap between observed and calculated discharge is very small during the 'dry' period in the autumn and the beginning of the winter but then starts increasing again until the wet period in the summer. VIC predicts the seasonality better, but seems to be a month early with the timing.

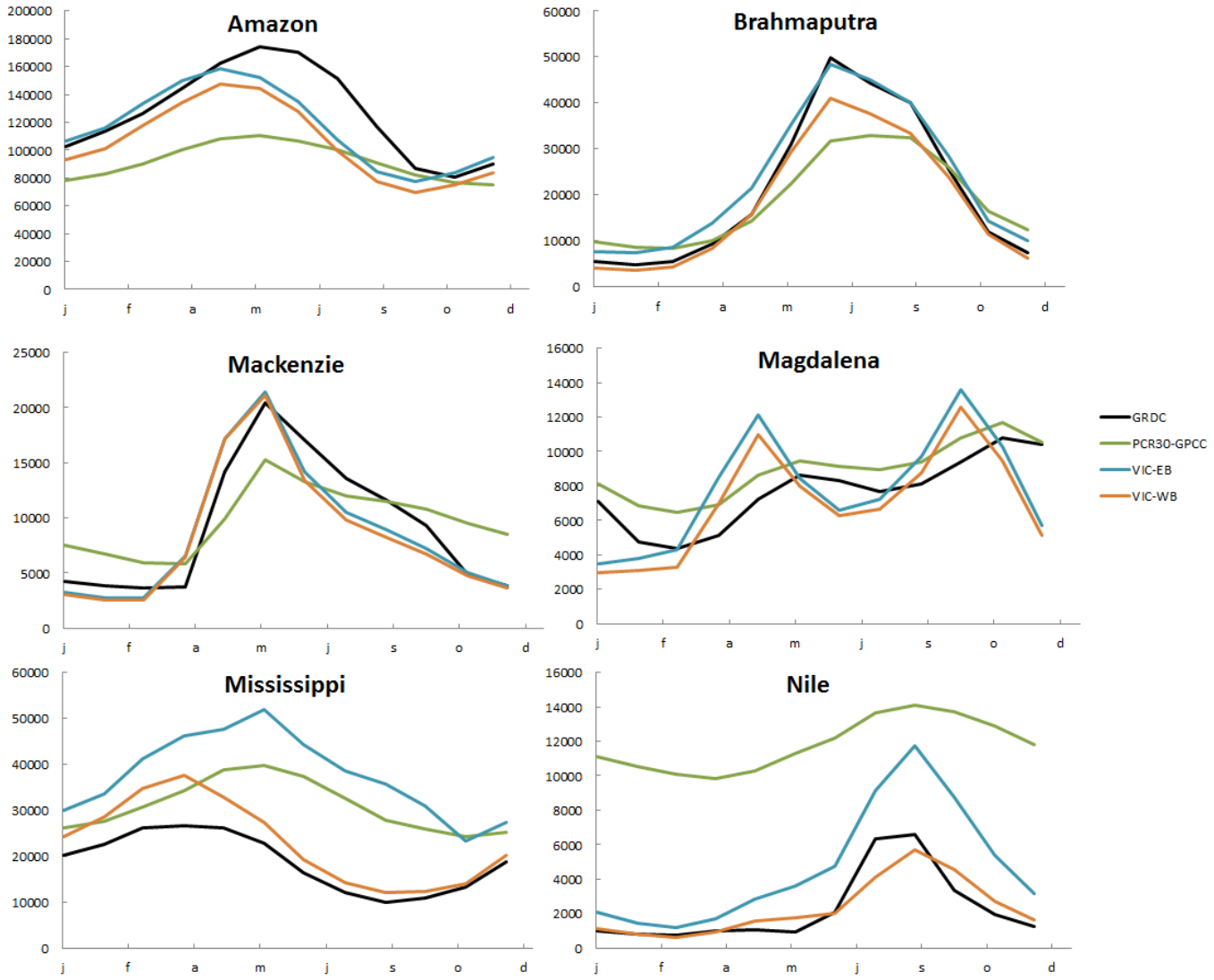


Figure 5: Discharge (in m^3/s) seasonality for the selected rivers for the period for which data is available from the GRDC. These periods are: for the Amazon (at Jatuares): 1992-2010, for The Brahmaputra (at Bahadurabad): 1985-2010, for the Mackenzie (at Arctic Red River): 1982-2010, for the Magdalena (at Calamar): 1980-1990, for the Mississippi (at Vicksburg): 1980-2010 and for the Nile (at Dongola): 1982-2002. PCR30-GPCC are the results of PCR-GLOBWB forced with the GPCC precipitation and with a resolution of 30 arcminutes.

The results show that the last two basins, the Mississippi and the Nile were the hardest to predict. For these, Only VIC-WB was successful. Especially the difference between VIC-EB and VIC-WB for the Mississippi is striking, as VIC has the best and the worst scoring model for this river. These two models perform similar to each other on the Amazon, Brahmaputra, Mackenzie and Magdalena but differ on the Nile and especially on the Mississippi. All models have trouble with the discharge during

the dry season for the Nile and the Mississippi. Especially with the Nile, PCR-GLOBWB overestimates the discharge for each month by a large amount. VIC-WB performs best for this river, whereas VIC-EB overestimates the peak discharge of the Nile, but predicts the dry season relatively better.

Table 4: Kling-Gupta, Nash-Sutcliffe efficiency scores and the coefficient of determination of the simulated discharge at an observation point for the different modes of the models. PCR-GLOBWB is indicated by PCR in the table. The same scores are also shown in Figure 6.

		Kling-Gupta efficiency			Nash-Sutcliffe efficiency			r^2		
Basin *(location)*	Model (mode)	PCR	VIC (EB)	VIC (WB)	PCR	VIC (EB)	VIC (WB)	PCR	VIC (EB)	VIC (WB)
Amazon *(Jatuares)*		0.36	0.74	0.72	-0.50	0.47	0.26	0.80	0.60	0.67
Brahmaputra *(Bahadurabad)*		0.56	0.85	0.78	0.74	0.92	0.90	0.88	0.94	0.94
Mackenzie *(Arctic Red River)*		0.53	0.89	0.85	0.62	0.80	0.76	0.71	0.82	0.79
Magdalena *(Calamar)*		0.69	0.26	0.33	0.57	-0.85	-0.64	0.89	0.25	0.29
Mississippi *(Vicksburg)*		0.35	-0.07	0.68	-1.58	-4.29	0.47	0.36	0.45	0.80
Nile *(Dongola)*		-3.23	-0.23	0.76	<-10	-0.88	0.72	0.61	0.81	0.72

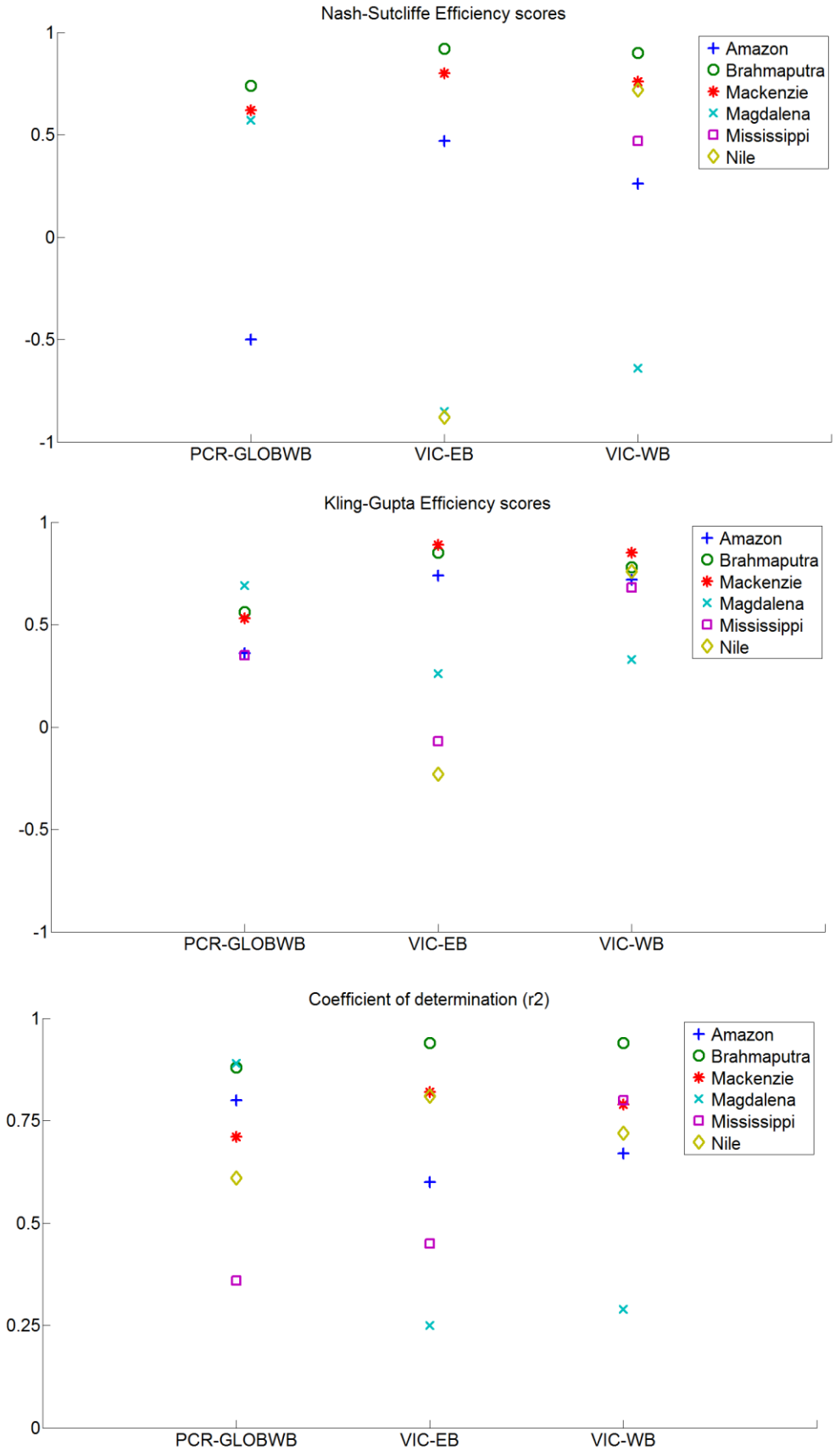


Figure 6: NSE, KGE en r^2 scores for the 6 basins. NSE scores for the Mississippi (-1.58) and the Nile (<-10) of PCR-GLOBWB are not included, as well as the NSE scores for the Mississippi (-4.29) of VIC-EB.

3.2. Runoff

The monthly runoff for each cell was estimated to be roughly equal to the precipitation subtracted by the calculated evapotranspiration rates. The monthly averaged difference between PCR-GLOBWB and VIC-EB is shown in Figure 7. It is visible that there are several regions in the world where the choice of model makes a difference in the amount of water available after evapotranspiration. The Amazon, Middle America, India and Southeast Asia have more water with VIC-EB than with PCR-GLOBWB, while for Indonesia, Europe, East Canada/US and Southeast America the opposite is true. Looking at the case study regions, both the Amazon and Brahmaputra have more water available for runoff in VIC, whereas the Southern Nile basin has more water in PCR-GLOBWB. The other three rivers do not show a difference between the models, as both models have areas in these basins where they have more water than the other model. This water availability has impact on the accumulated discharge in the rivers. Both the Amazon and the Brahmaputra have a higher average calculated discharge in VIC-EB than in PCR-GLOBWB (Figure 5), while the Nile has a higher discharge calculated by PCR-GLOBWB.

Despite the fact that there are several areas where one of the models has more water available than the other model, the correlation between both models is good (Figure 8). The majority of cells has a correlation score above 0.5, the correlation only fades away to scores below 0.5 in small areas in Central Canada and the Sahara. Some of the highest scores (above 0.95) occur in the Amazon basin and Southeast Asia, the areas where the average monthly differences are the highest.

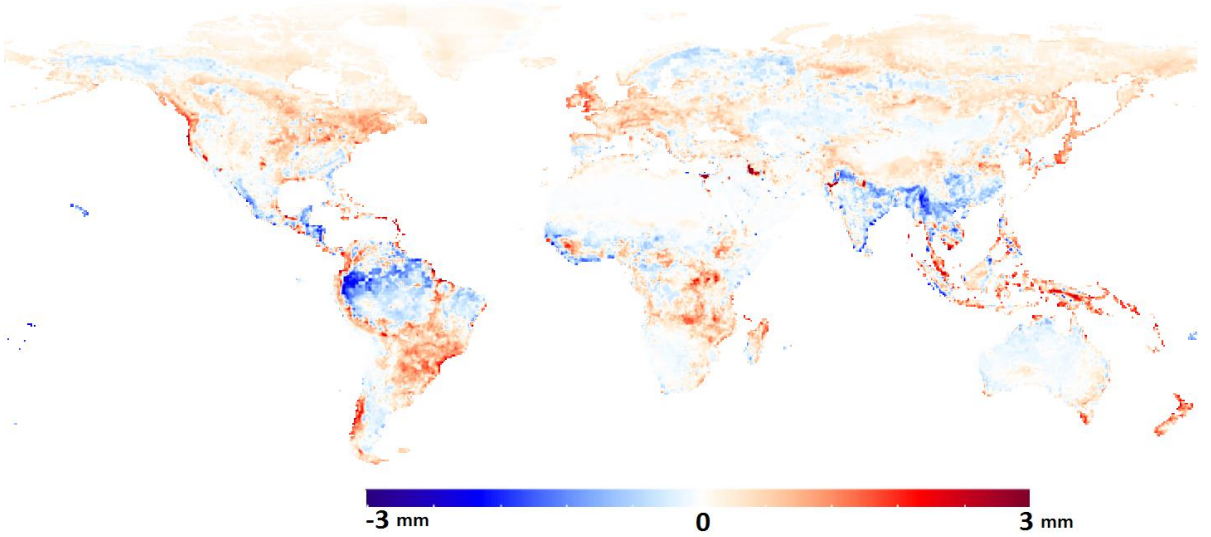


Figure 7: Difference of the monthly averaged precipitation minus evapotranspiration rates as calculated by PCR-GLOBWB and VIC-EB (in mm/day) and averaged for the period 1979-2010. Positive values mean that the average daily precipitation minus evapotranspiration is larger in PCR-GLOBWB than in VIC-EB and vice versa. Note that this does not show accuracy, only the difference between the amount of water available after evapotranspiration for both cells. Scores above 3 mm occur in Egypt, Iraq/Iran and some coastal cells (in the range 4-6mm).

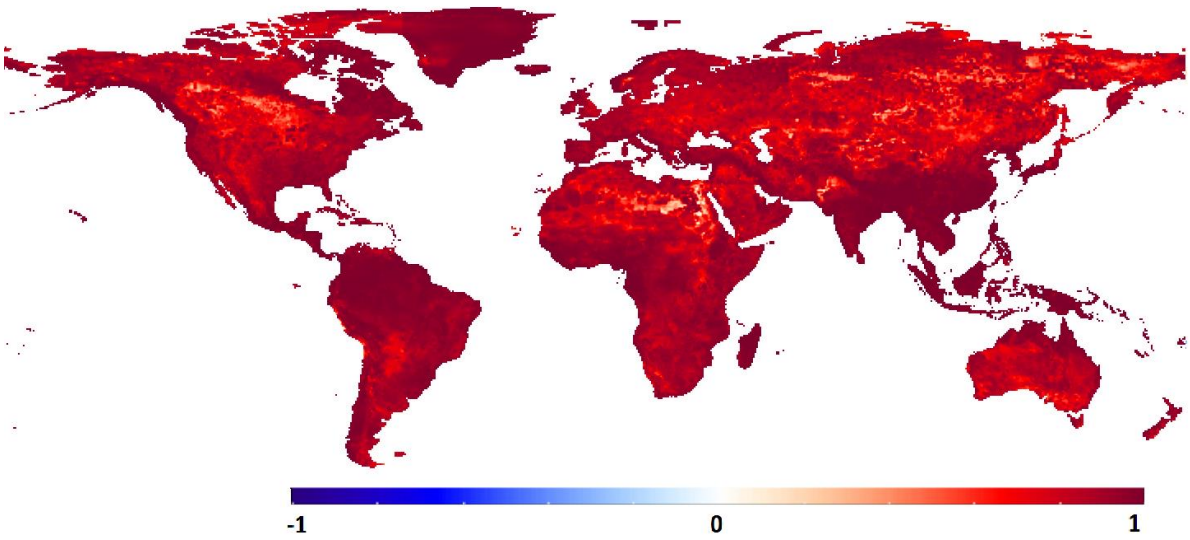


Figure 8: Correlation between the monthly averaged precipitation minus evaporation rates of PCR-GLOBWB and VIC-EB for the period 1979-2010.

Figure B1 (in Appendix 2) shows the difference of the average monthly precipitation minus evapotranspiration rates between VIC-EB and VIC-WB. In contrast to the difference between PCR-GLOBWB and VIC-EB, where both models had areas where they had the most water available, Figure B1 shows that VIC-EB has more water available than VIC-WB for almost the entire world. This is also shown in the discharge graphs (Figure 5), as the amount of water in the channel calculated by VIC-EB is higher than that of VIC-WB. The correlation between the VIC models is on the whole good (Figure B2) with scores above 0.9 for almost the entire world.

3.3. Snow water equivalent (SWE)

The SWE comparison was done by comparing the monthly averaged SWE results of VIC and PCR-GLOBWB to the ASMR-E SWE data product. The pattern of the snow simulations of all three models is similar, which is shown in Figure 9 and Table 5. All three comparison criteria are close to each other. PCR-GLOBWB has the highest scores for KGE and NS, but the margin with VIC is small. Both VIC models have a higher correlation with the ASMR-E data, but only by a small margin. VIC-EB scores slightly better than VIC-WB, but the difference is smaller than the difference with PCR-GLOBWB. The median of the KGE and the NS are below zero, indicating that the models failed to achieve the 'better than the mean of the observations' benchmark. However, this is the median for the whole world. Figure 9 shows that there are several regions where the KGE and NS scores are positive and thus perform better than the mean of the observations in those cells. On the other hand, there are also several regions where the scores are negative. The regions with the highest scores are located in Eastern Europe, inland Scandinavia and central Canada. The lowest scores occur along the Atlantic coast (on both sides), the American Pacific coast and central Asia. For most areas, the VIC models score more positive than PCR-GLOBWB when the scores are positive, but also more negative when the scores are negative. The column 'ASMR-E only +' in Table 5 confirms that the median of only the positive scores of VIC is higher than PCR-GLOBWB.

The correlation with the ASMR-E data is quite strong, as there are only a few cells with scores below 0.5. Only the southern parts, where the occurrence of snow was limited to only a few months in the study period, score lower. The pattern of high and low correlation scores is similar between models as well, but it is not the same pattern as for the KGE and NS scores.

Both VIC models have more extreme scores for all comparison criteria. Because both extremes are more pronounced in VIC, the median remains similar to PCR-GLOBWB, but the range is larger. This is also visible in the boxplots in Figure 10. The median is at the same level, but both the interquartile range and the lower whiskers are larger, indicating a larger spread in the calculated SWE values.

Table 5: Median KGE, NS and correlation scores for the Snow Water Equivalent (SWE) output of all three models (PCR-GLOBWB, VIC-EB and VIC-WB). Compared to ASMR-E SWE data and to the output of PCR-GLOBWB (indicated by PCR in the table). 'ASMR-E only +' is the median score of only the positive efficiency scores for that particular score.

	Kling-Gupta efficiency			Nash-Sutcliffe efficiency			Correlation		
	ASMR-E	ASMR-E only +	PCR	ASMR-E	ASMR-E only +	PCR	ASMR-E	ASMR-E only +	PCR
PCR-GLOBWB	-0.033	0.319	1	-0.094	0.313	1	0.692	0.756	1
VIC-EB	-0.055	0.341	0.737	-0.141	0.335	0.802	0.699	0.770	0.941
VIC-WB	-0.058	0.341	0.739	-0.143	0.336	0.802	0.696	0.770	0.941

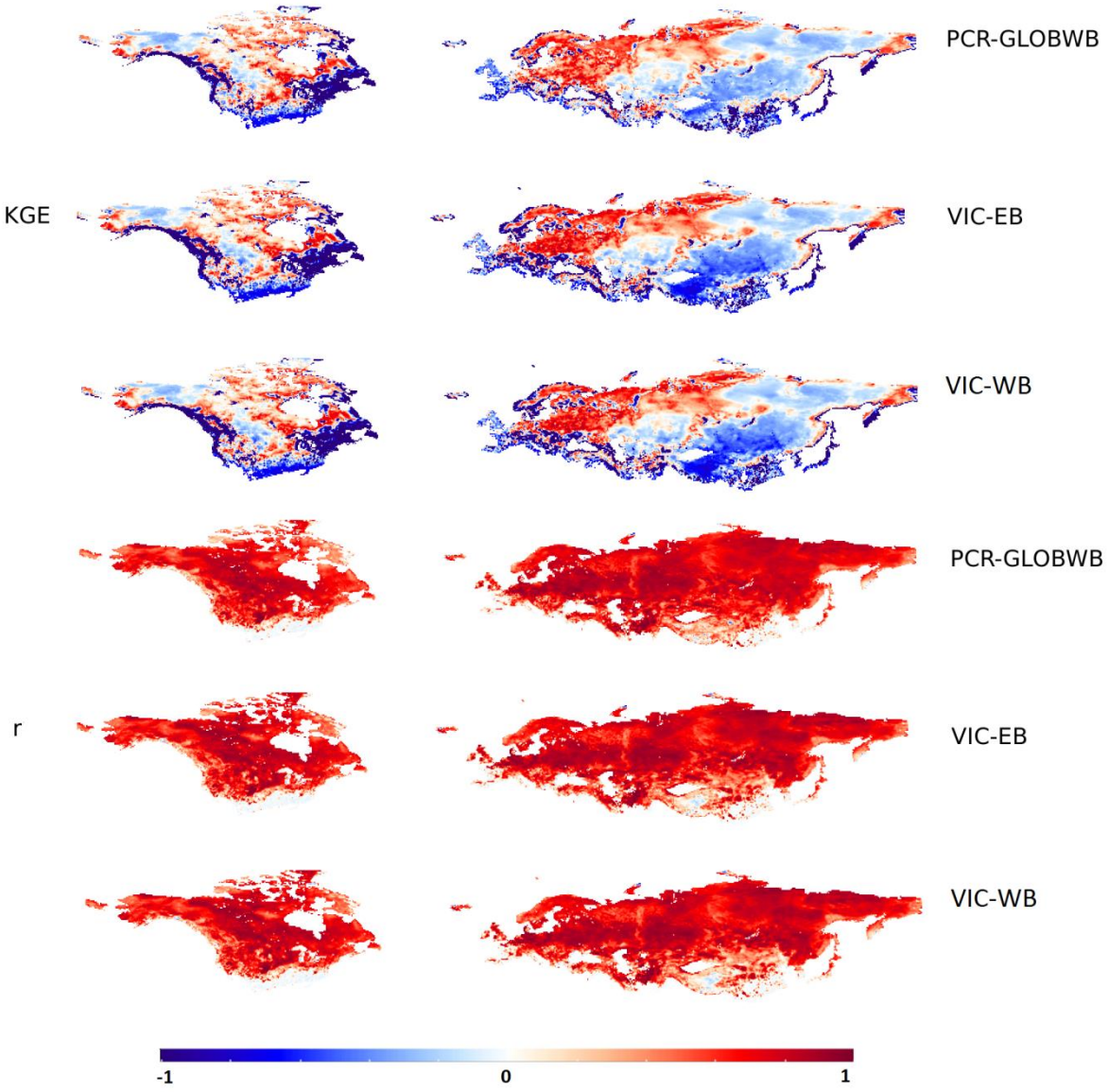


Figure 9: KGE and correlation scores of Snow Water Equivalent (SWE) output of PCR-GLOBWB, VIC-EB and VIC-WB when compared to ASMR-E SWE data.

The output of the models is also compared to the SWE output of PCR-GLOBWB instead of the ASMR-E SWE. The results are shown in Figure 11 and Table 5. It shows that the results of VIC and PCR-GLOBWB are similar for large parts of the study area. The KGE and NS scores are high for almost the entire Northern Hemisphere. The only negative areas are located in Western Europe and some spots in the Rocky Mountains, central China and the Tibetan Plateau. The correlation plots indicate that the temporal patterns in both models are comparable, with correlation scores above 0.90 for almost the entire area. The areas with the largest KGE and NS score difference in the ASMR-E comparison also have the lowest correlation scores (Figure 9), which indicates that the VIC and PCR-GLOBWB outputs differ at these locations.

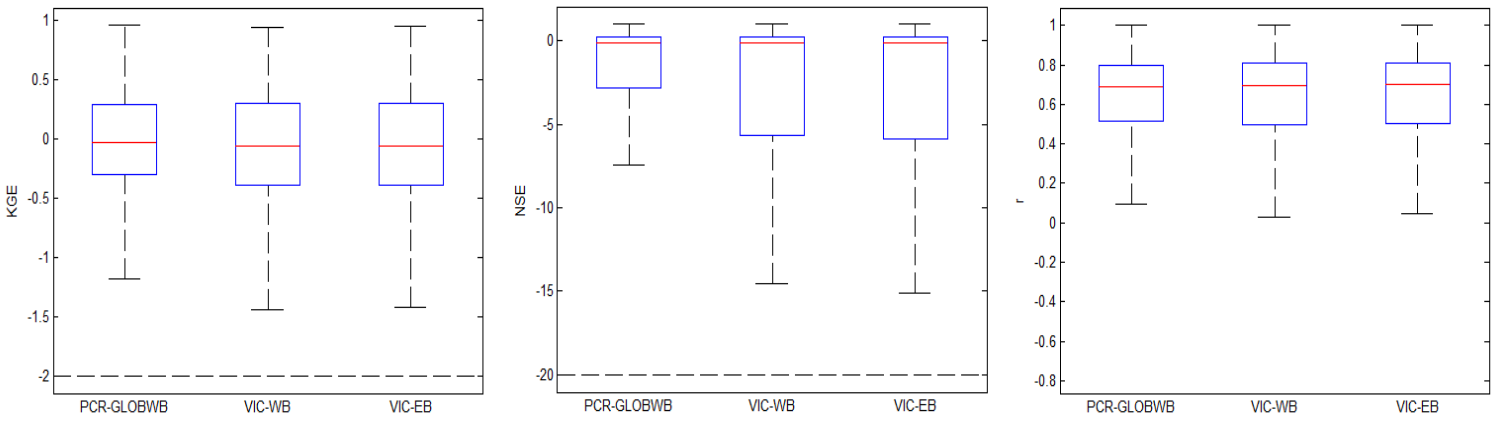


Figure 10: From left to right: KGE, NS and correlation scores for the global runs of PCR-GLOBWB, VIC-EB and VIC-WB compared to ASMR-E Snow Water Equivalent data. Outliers are not shown in the boxplots.

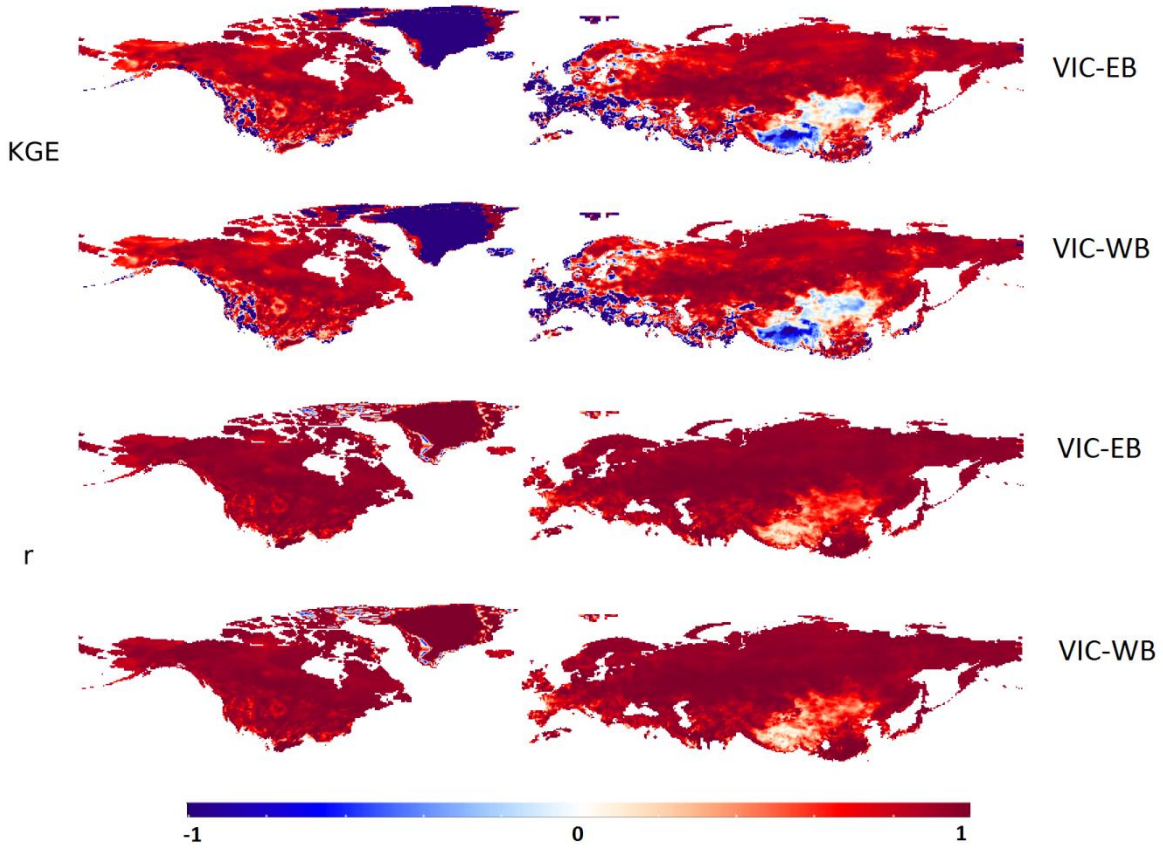


Figure 11: KGE and r scores for both VIC models when compared to the Snow Water Equivalent output of PCR-GLOBWB.

3.4. Evapotranspiration

Output to Fluxnet

This comparison was done by comparing the monthly averaged ET results of VIC and PCR-GLOBWB to the Fluxnet ET data product. Looking at the spatial distribution of the KGE and NSE scores of both models with the Fluxnet dataset, the first observation is the occurrence of a belt of low (<0) scores along the tropics (NSE shown in Figure 12). Large parts of the Amazon, the Caribbean, tropical Africa, Indonesia and Northwestern Australia score lower than the other regions in both hydrological models. For the whole world, the median scores in Table 6 indicate that PCR-GLOBWB has the highest median for the KGE (0.464), while the VIC models score lower (median of 0.335 for EB and 0.236 for WB). The values of VIC are also more skewed to the lower values than PCR-GLOBWB (Figure 14). This is also visible in the spatial plots. Even though both models show the blue belt of low KGE scores along the tropics, the belt of PCR-GLOBWB is smaller. Furthermore, the scores in Europe and the US are also higher than for VIC-EB. The NSE shows the same pattern, PCR-GLOBWB has a higher median and more positive scoring areas than VIC-EB, which scores better than VIC-WB.

The correlation scores do not show the same belt, there are only low (negative) scores for the Amazon and to a lesser extent in central Africa and Indonesia (Figure 13). Apart from these three areas, the correlation scores are generally high, with the highest scores occurring in the northern hemisphere. Except for Northern Africa and the Middle East, the scores rarely dip below the 0.8 for both VIC models and PCR-GLOBWB has mostly correlation scores above 0.9 in this region. Northern Africa and the Middle East remain positive, but the correlation is weaker than in the other regions, as both models score in the 0.2-0.5 range for these regions.

When comparing VIC-WB to VIC-EB, the blue belt along the tropics is even more pronounced in VIC-WB. It stretches further south and southern Africa and Argentina score visibly lower as well. Also in other parts of the world it can be seen that VIC-WB has lower KGE values. This is also shown by the boxplot.

Output to ERA-Interim

A similar pattern as with the comparisons with Fluxnet is visible when the output of the models is compared to the monthly ERA-Interim ET data (Figure 12 and Figure 13). All models have a large belt of lower scores along the tropics, but the belt is less pronounced and especially Eastern Africa, the Middle East and Australia score higher. The global boxplots show that for VIC, the KGE, NS and correlation scores are all higher than with the Fluxnet comparison (Figure 14). Even though both the KGE and NS scores have improved, the correlation scores only improve marginally. The boxplots also show that the spread of the correlation has increased.

Table 6: Median KGE, NS and correlation scores for the ET output of all three models (PCR-GLOBWB, VIC-EB and VIC-WB), compared to FLUXNET-MTE data, ERA-Interim data and to the output of PCR-GLOBWB (indicated by PCR in the table).

	Kling-Gupta efficiency			Nash-Sutcliffe efficiency			Correlation		
	Flux	Era	PCR	Flux	Era	PCR	Flux	Era	PCR
PCR-GLOBWB	0.464	0.437	1	0.408	0.320	1	0.895	0.895	1
VIC-EB	0.335	0.403	0.242	0.005	0.156	-0.175	0.826	0.829	0.860
VIC-WB	0.236	0.354	0.113	-0.197	0.078	-0.454	0.813	0.818	0.848

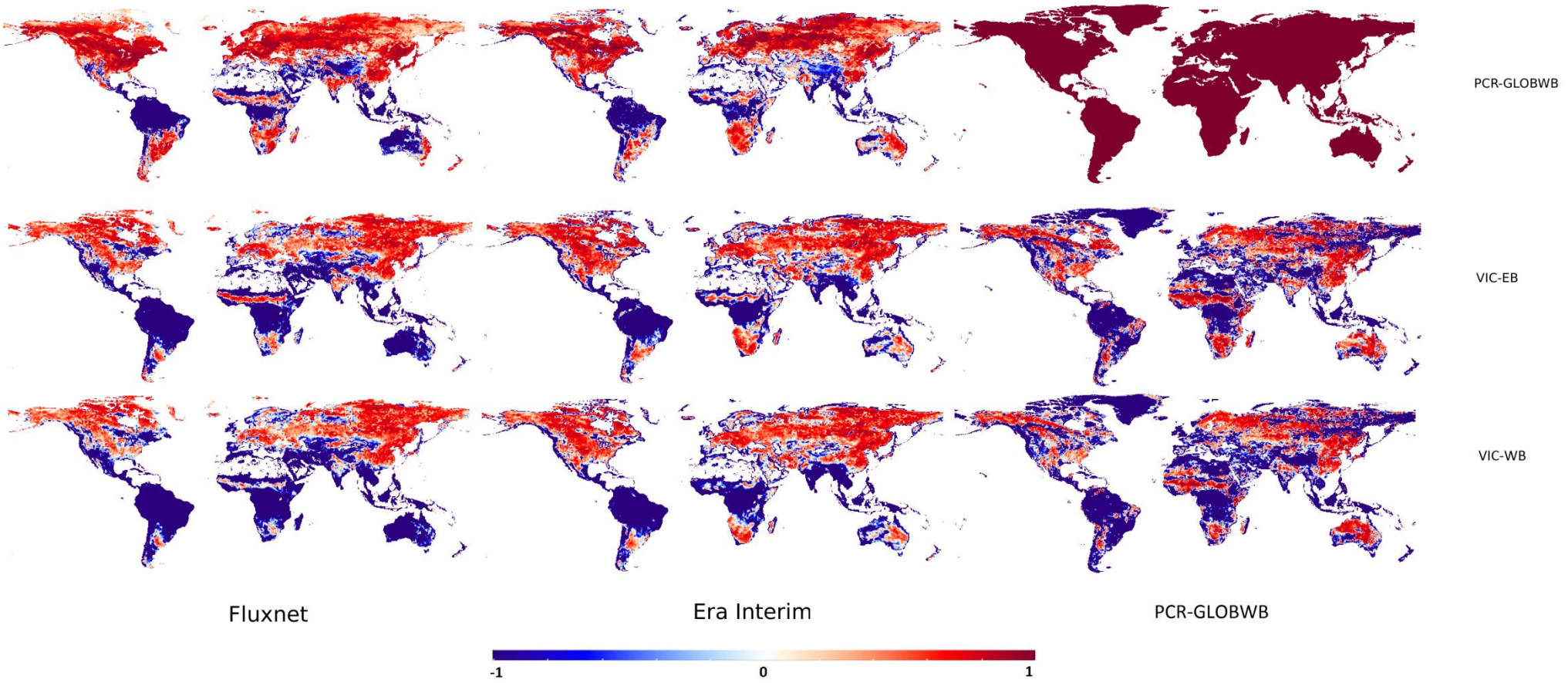


Figure 12: Nash-Sutcliffe efficiency scores for the ET model output of the three models (PCR-GLOBWB, VIC-EB and VIC-WB) compared to the three ET datasets (Fluxnet-MTE, ERA-Interim and the output of PCR-GLOBWB).

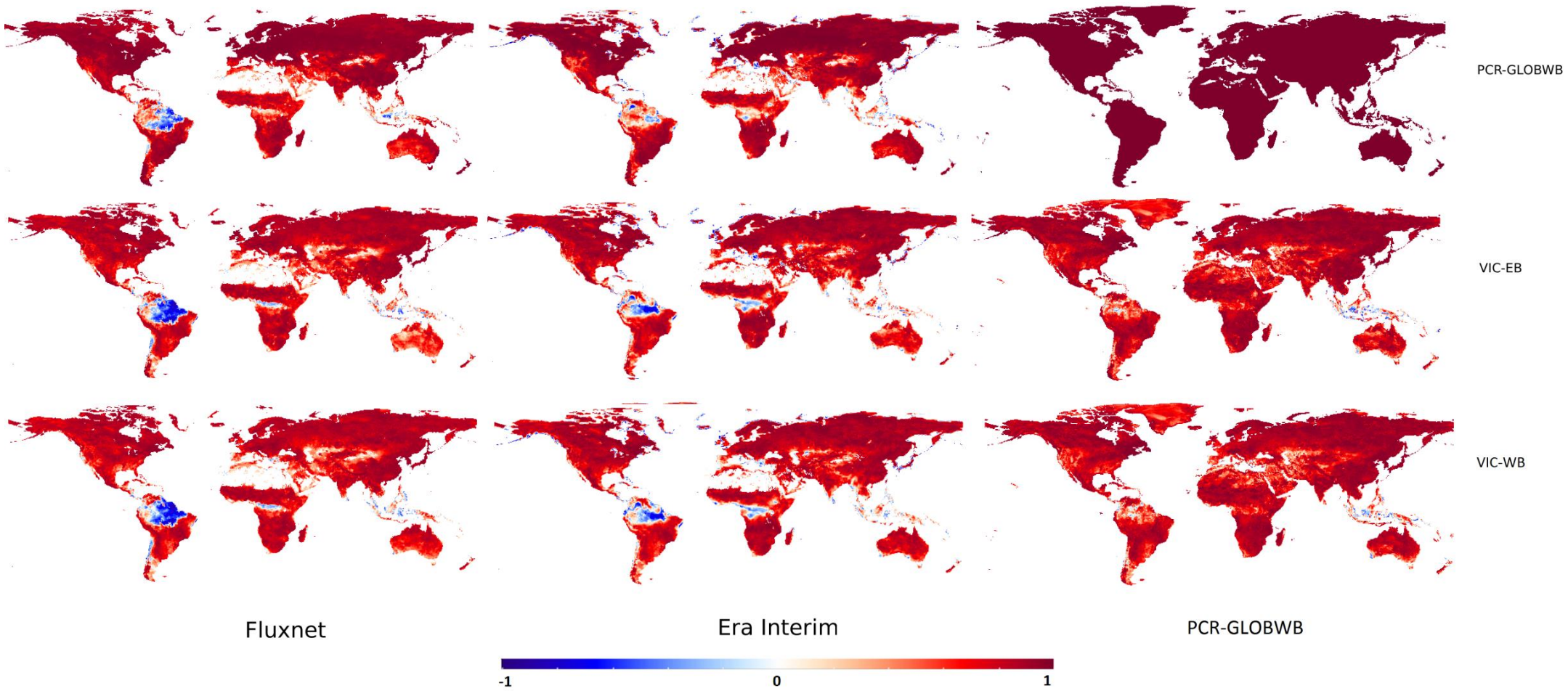


Figure 13: Correlation scores for the ET model output of the three models (PCR-GLOBWB, VIC-EB and VIC-WB) compared to the three ET datasets (Fluxnet-MTE, ERA-Interim and the output of PCR-GLOBWB).

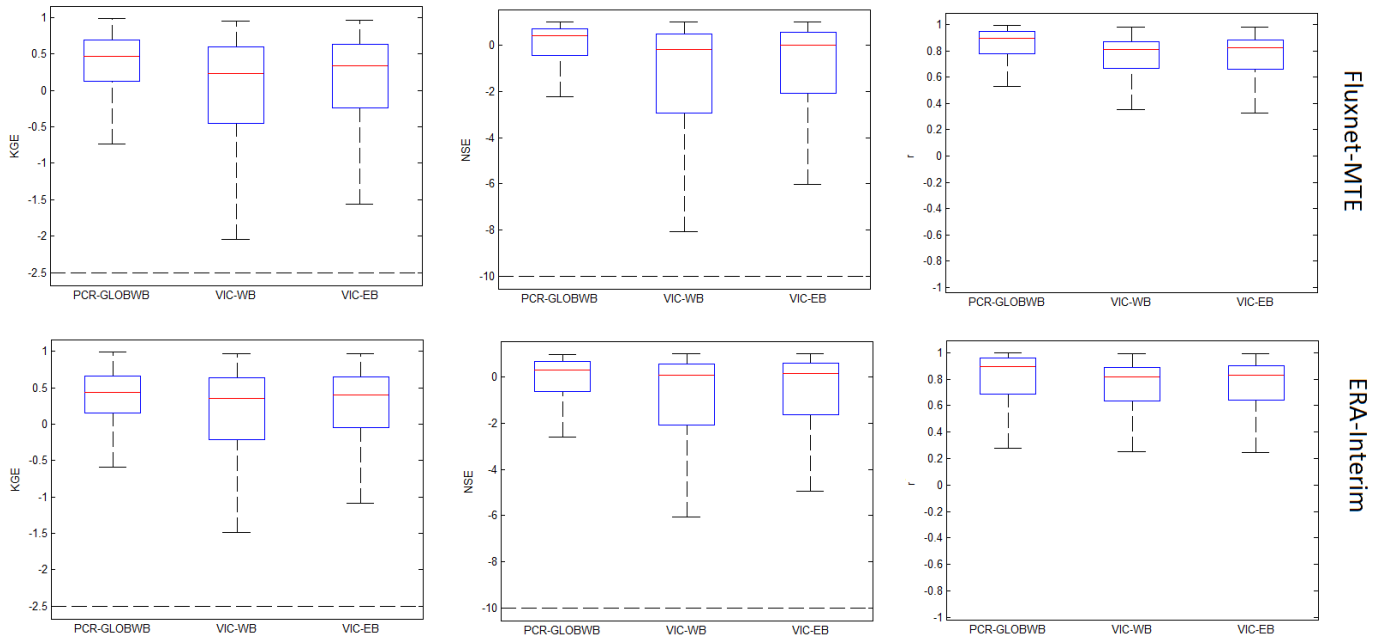


Figure 14: From left to right: KGE, NS and correlation scores for the global runs of PCR-GLOBWB and VIC compared with ERA-Interim and Fluxnet comparison dataset. Outliers are not shown in the boxplots.

The boxplot scores for PCR-GLOBWB decreased in the ERA-Interim comparison, but they remain slightly higher than the scores for the VIC models (Figure 14 and Table 6). There are also areas in the world where both models score worse when comparing the data to ERA-Interim. Examples are India and the Sahel region in Africa. For the VIC models the decrease in accuracy over India and the Sahel is countered by the increase in South America, Southern Africa and Australia, which leads to higher global scores. For PCR-GLOBWB the increase is not sufficient (Southern Africa and Australia) or non-existent (South America) to counter the decrease in other areas and therefore the global scores decrease.

Important to note is the fact that PCR-GLOBWB obtains the highest comparison scores with the Fluxnet dataset, while the VIC models perform better when their output is compared to the ERA-Interim ET data.

With ERA-Interim, there also appears to be an issue with the coastline. For all three models, grid cells that have a high fraction of water (either coastal cells, or cells containing large lakes), have a lower KGE or NS score compared to neighboring cells that have a lower fraction of water. There is a similar pattern for the correlation, but it is much less pronounced than for the KGE and NS scores.

Output to PCR-GLOBWB

When comparing the output of VIC directly to the output of PCR-GLOBWB, it becomes clear that there are large differences between the results of the models for several areas around the world, but also several areas where the models perform similarly. The NSE and KGE scores are positive for large areas in China, Russia, Australia, North America and the Sahel in Africa, indicating that there is similarity in the results for these areas (Figure 12). However, in the tropical zone and Europe the scores are negative, indicating differences between both models.

The correlations between VIC and PCR-GLOBWB are comparable to the correlation scores between VIC and ERA-Interim or Fluxnet. The median scores of the VIC-PCR-GLOBWB correlation are slightly lower than those of ERA-Interim or Fluxnet (Table 6), but there are less areas with a negative correlation. Especially the Amazon, which has low correlation scores for ERA-Interim and Fluxnet for

both models, has mostly positive correlation values when the models are compared to each other (Figure 13). However, despite the positive correlation, the KGE and NS values for both VIC-EB and VIC-WB for this area are below zero. This means that there are substantial differences between both models in this area, even though the pattern of the output is similar.

3.5. Soil Moisture

The soil moisture output was compared to the monthly volumetric soil moisture data from the ESA CCI dataset (Figure 15). For large parts of the world the KGE and NS scores are similar for VIC and PCR-GLOBWB. There are some differences between both models as well. PCR scores better for parts of central Russia, Northeastern Canada and Australia. But VIC has better KGE and NS scores for most of China and the Himalayas, as well as for the Sahara desert, the Arabian Peninsula and large parts of South America. Areas where PCR-GLOBWB scores poorly are either dry or wet. This also means that VIC scores better for some of the study areas. In the Mississippi, Amazon and the Egyptian Nile basin the scores for both VIC models are higher than the PCR-GLOBWB scores. Given the fact that the upper soil layer plays a large role in the generation of runoff during rain events, this could have an effect for the peak discharge of the rivers in these regions. The boxplots show that PCR has lower values than VIC for all three categories. The low medians and high spread indicate that there is a large spatial difference for the models. This is also visible in the spatial plots. There are large areas where the KGE and NS is low (<0) for both models. Especially in the north, all three comparison criteria are very low. Apart from the north, the correlation is only low in the Sahara and some areas in the tropics.

The differences between VIC-EB and VIC-WB are very small. The median KGE and NSE scores for VIC-WB are slightly higher than for VIC-EB (Figure 15 and Table 7), but the differences are small and no clear spatial pattern can be seen in the distribution of the differences. The correlation differences between both models are even lower. All put together, VIC-WB scores slightly better than VIC-EB for the KGE and NS, with the same score for the correlation.

Figure B3 shows the KGE and correlation scores when VIC-EB and VIC-WB are compared to the PCR-GLOBWB output. It confirms the differences in model results that are visible in Figure 15 and shows that for the majority of the world the results of VIC and PCR-GLOBWB are relatively close to each other (KGE score > 0.3). The correlation between the models is even better. Almost all areas show the same pattern in their output. The Nile in Egypt and some patches in northern Russia and Canada are lower (0.0-0.1), but only a small amount of the calculated gridcells has a negative correlation.

Table 7: Median KGE, NS and correlation scores for the soil moisture output of all three models compared to ESA CCI soil moisture and to the output of PCR-GLOBWB (indicated by PCR in the table).

	Kling-Gupta efficiency		Nash-Sutcliffe efficiency		Correlation	
	ESA	PCR	ESA	PCR	ESA	PCR
PCR-GLOBWB	0.036	1	-2.16	1	0.40	1
VIC-EB	0.10	0.28	-1.86	-3.22	0.47	0.72
VIC-WB	0.11	0.28	-1.70	-3.23	0.47	0.72

Since the calibration of VIC has influence on the infiltration and runoff of the model, it is possible that the soil moisture scores are different in calibrated and uncalibrated areas. Figure 16 shows boxplots with the comparison scores of the Amazon (calibrated in VIC) and the Congo basin (uncalibrated). The boxplots indicate small differences between the scores. The largest difference occurs for the KGE score of PCR-GLOBWB, where the Congo basin has a higher score than the Amazon. The median scores for VIC are similar to each other, there are improvements for the KGE and correlation scores in the lower quartile, but the best 50% of the scores are similar for the Amazon and the Congo.

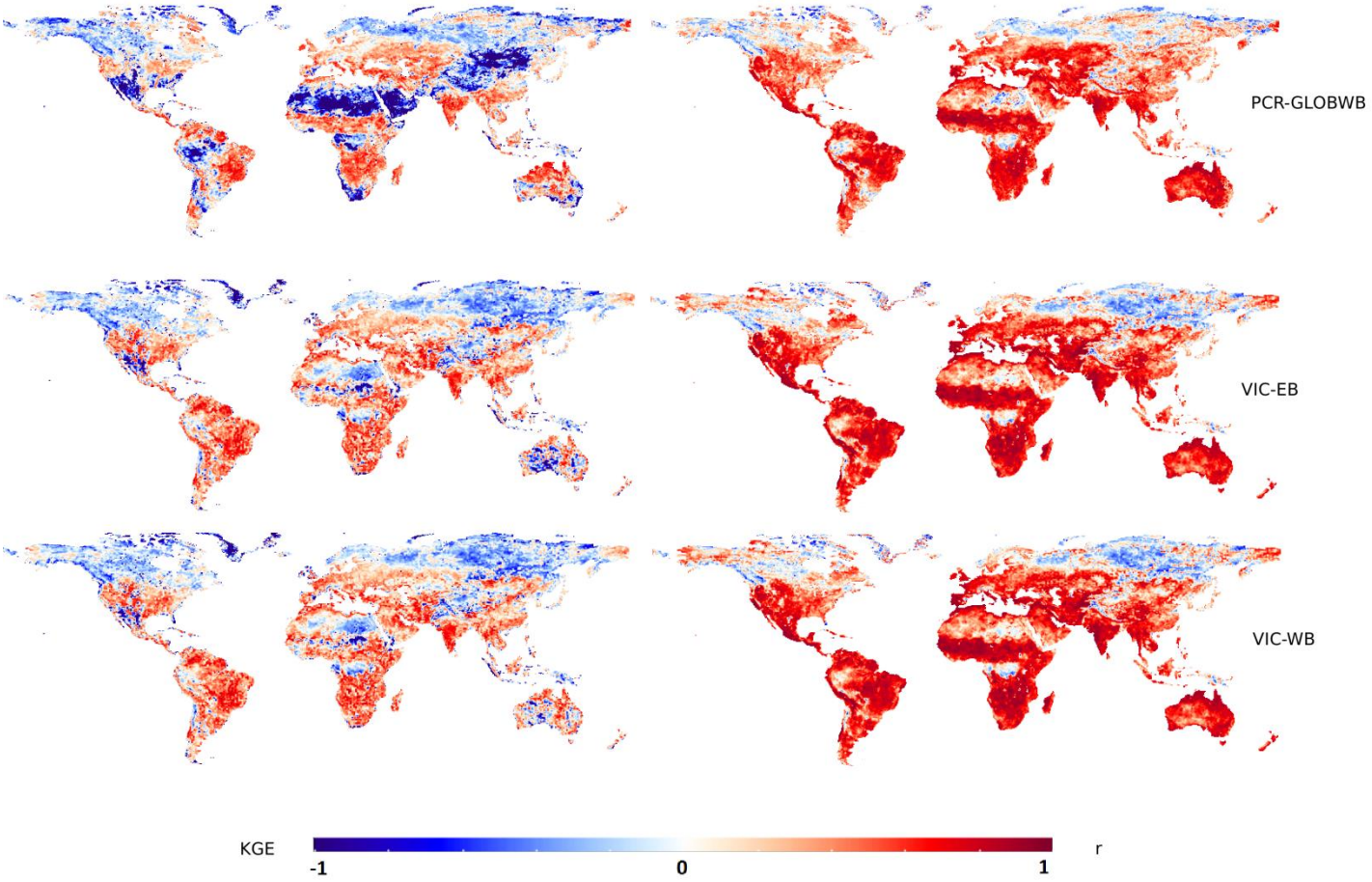


Figure 15: KGE efficiency and correlation scores of soil moisture output of PCR-GLOBWB and both VIC models compared to the ESA CCI dataset.

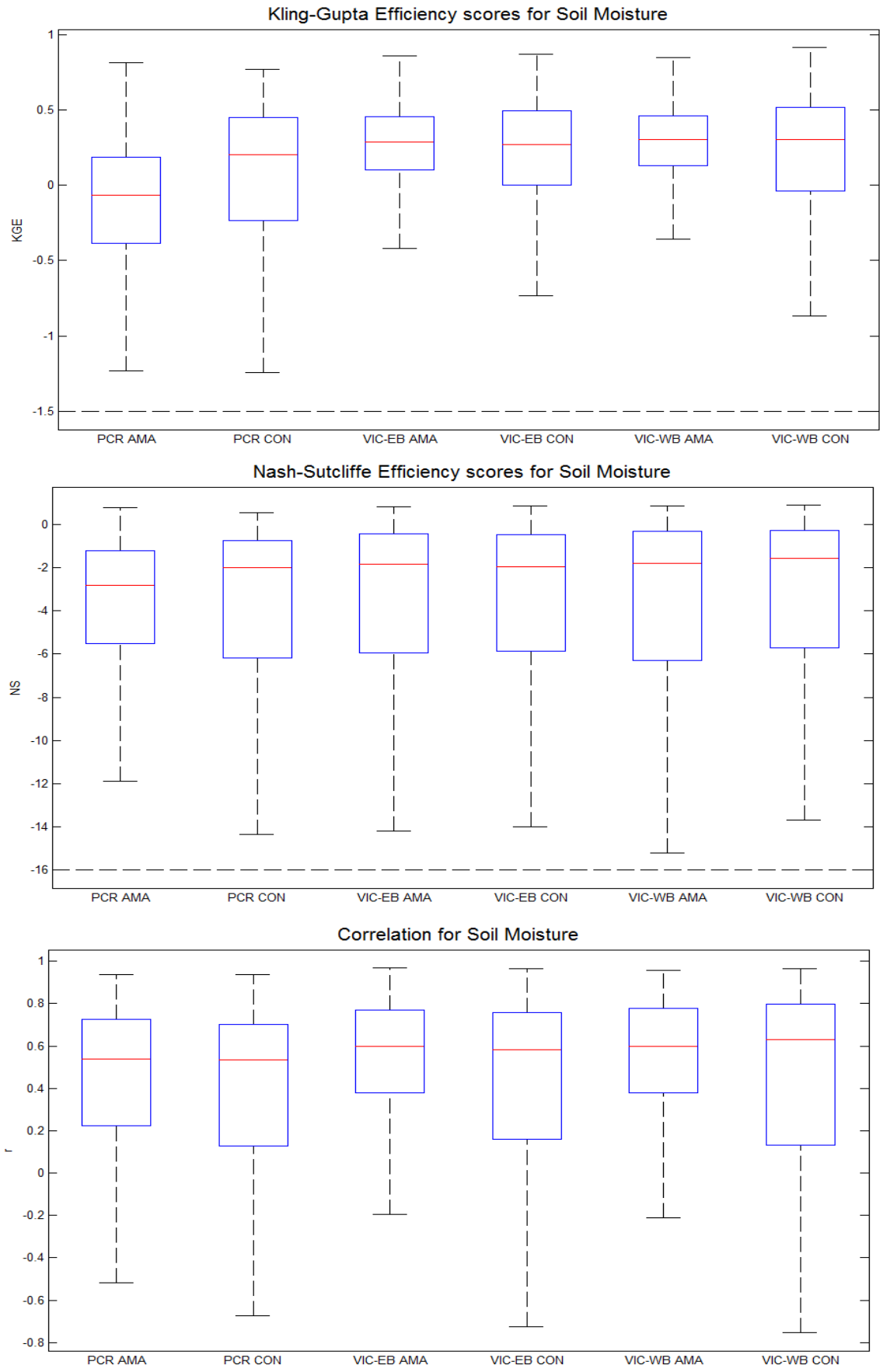


Figure 16: Top to bottom: KGE, NSE and correlation scores for output of the models (PCR is PCR-GLOBWB) in the Amazon (AMA) and the Congo (CON) basin compared to ESA CCI soil moisture data. Outliers not shown in the figure.

4. Discussion

In the ideal situation, the datasets used to compare to the results of VIC and PCR-GLOBWB are a perfect representation of reality. However, each of the reference dataset has its own shortcomings and regions and periods where the accuracy is lower. This makes the evaluation of the model differences more difficult, as not only the model outcomes need to be taken into account, but also the reliability of the comparison dataset. The evaluation was done by answering the research questions. The sensitivity of PCR-GLOBWB to changes in climate input, resolution and evaporation method will be discussed in Appendix 1.

An indicator of the accuracy of a dataset is the correlation with the output of the models. When certain cells have high correlations for both VIC and PCR-GLOBWB, the chance that the control dataset has realistic values increases. Both models give a calculated estimation of the water balance in a certain cell, based on external forcing's. If both of them give a certain pattern for a cell, this is based on those external forcing's and the internal parameters of that cell. However, the models do not use the same internal parameters for each cell. Thus, if both models predict a certain pattern, there is a high chance that this pattern is the same for the reality. Therefore, one of the first indications of the quality of the comparison data is the correlation between both models and the comparison dataset. On the other hand, when the comparison data has a negative correlation with both VIC and PCR-GLOBWB, the pattern found by the comparison data is the reversed pattern of both VIC and PCR-GLOBWB. This raises doubt about the quality of the comparison data or the output of the models for that area.

It has to be taken in consideration that most of the comparison was done with monthly averaged data. This was done because the used remote sensing data also has a monthly resolution. By using monthly averaged data the extreme values become less pronounced. This leads to more subdued results and inherently higher scores than if the comparison was done on a daily resolution. However, as there are still differences between the model output and the remote sensing, it remains acceptable to use this method for the evaluation.

4.1. What is the effect of the energy balance in a global hydrological model on the discharge of a major river and on local hydrological variables, such as evapotranspiration, snow water equivalent and soil moisture?

The discharge graphs show that the predicted discharge between PCR-GLOBWB and the VIC models is different. Important to note is also that, in five of the six studied discharges, a VIC model has the highest KGE and NS scores. VIC-EB scores the highest for the Amazon, the Brahmaputra and the Mackenzie and VIC-WB for the Mississippi and the Nile. Important to note is that VIC-WB performs similar to VIC-EB for the Amazon, Brahmaputra and Mackenzie (where VIC-EB scores the highest), but VIC-EB scores worse for the Mississippi and the Nile (where VIC-WB scores highest). The Magdalena is the only river where PCR-GLOBWB has the highest efficiency scores. This corresponds to the basins that were calibrated in VIC, only the Magdalena was not calibrated and this basin has the lowest scores. Calibration has a large effect on the discharge predictions of a model and it is probable that this is the reason for the more accurate discharge results of VIC. However, as the discharge is an accumulation of the runoff generated on a cell level, the local hydrology (the ET, soil moisture and SWE) plays an important part as well.

As explained in the introduction, the expectation was that the model with the energy balance (VIC-EB) would have an advantage over the other models in the areas where energy based processes (ET and snow melt etc.) play a large role. However, the results do not correspond with this hypothesis. The ET output of PCR-GLOBWB corresponds better to both the Fluxnet and the ERA-Interim ET comparison data than the output of either VIC model. Moreover, PCR-GLOBWB shows higher scores on a cell to cell basis for the tropics, an area where the amount of ET plays an important role and higher scores of VIC were expected.

VIC-EB is also compared to VIC-WB to explore the added value of an energy balance, as the energy balance is the only difference between VIC-EB and VIC-WB. The results are in favor of the VIC-EB. It has higher median values for all three categories with both ET comparison datasets and the boxplots and spatial maps show a better spread of the scores as well. Based on this, it can be concluded that an energy balance has a positive effect on the accuracy of the ET. However, other elements in the model structure of PCR-GLOBWB and VIC have influence on the ET as well and this leads to higher scores for PCR-GLOBWB.

The snow module is also dependent on the energy balance in VIC. Especially since PCR-GLOBWB has a different method to model snow; differences between the three models were expected in the snow results. However, this is not the case. No clear differences were observed between the three models and the accuracy scores are very high when the VIC models are compared to PCR-GLOBWB.

The high degree of similarity between the VIC and PCR-GLOBWB SWE output raises doubts about the accuracy of the ASMR-E SWE products. The high correlation scores between both VIC and PCR-GLOBWB with the ASMR-E comparison dataset indicate that the snow pattern is included accurately in all three datasets (VIC, PCR-GLOBWB and ASMR-E). However, the low KGE and NS scores for the same areas suggest that there is a difference between the SWE obtained by ASMR-E and the SWE calculated by VIC and PCR-GLOBWB. Given the fact that both models perform very similar to each other for extensive areas, it would appear that ASMR-E is the weakest link in the chain. A different SWE dataset could be used to test this.

Gao et al. (2010b) tested the accuracy of the AMSR-E snow products for a test area in central Alaska. In this test area an accuracy of 68.5% was obtained and the main issue was an overestimation of the SWE values by ASMR-E. Figure 17 shows the mean monthly difference between ASMR-E SWE values and the output of VIC-EB and PCR-GLOBWB. Based on the equations of the KGE and NS, in cells where the mean difference is large, the KGE and NS scores are low. This figure also shows that the difference between ASMR-E and the models is positive above Alaska, indicating that the output of

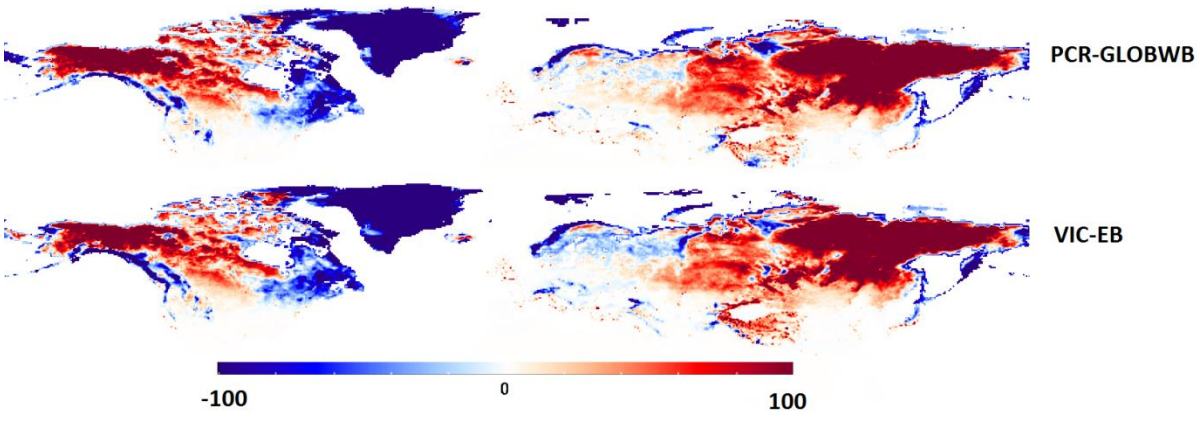


Figure 17: Mean monthly difference between SWE registered by ASMR-E and by PCR-GLOBWB and VIC-EB in mm. Positive values indicate that the ASMR-E dataset has a larger SWE value for that cell than PCR-GLOBWB or VIC-EB.

ASMR-E is larger than the output of the models. This corresponds to the overestimation found by Gao et al. (2010b). However, other areas show the opposite: parts of Russia, Northern Europe, East Canada and the coast of western Canada show that the VIC-EB and PCR-GLOBWB simulations are higher than the ASMR-E observations, but the difference is generally smaller than for the areas where the opposite is true. Examining the difference over time shows that ASMR-E predicts higher SWE values than VIC and PCR-GLOBWB during the winter, but lower during the spring. This indicates that there is also a discrepancy between the melt seasons observed by ASMR-E and simulated by PCR-GLOBWB and VIC.

Because the results of VIC and PCR-GLOBWB are so similar, it is impossible to draw decisive conclusions about the benefit of an energy balance for snow simulations where the amount of snow on a seasonal basis is an important issue. The differences between both models are very small and could easily be caused by the different snow modules of the models. It is interesting to see that, despite its relatively simple degree day snow module compared to the more sophisticated VIC snow module, PCR-GLOBWB scores similar to both VIC models. However, VIC has both higher and lower KGE/NS values than PCR-GLOBWB. This indicates that it scores better than PCR-GLOBWB in the areas where both models have positive KGE/NS scores, but also worse for the areas where the simulations differ more from the observations. As explained before, this could be due to the accuracy of the ASMR-E SWE observations. If only the positive comparison scores are taken into account, the VIC models compare better to the ASMR-E data than PCR-GLOBWB. This difference is due to the different snow module, of which the energy balance is an important part.

Despite the fact that VIC-WB uses the energy balance in the snow module, the VIC-EB results give slightly higher median scores for the KGE, NS and the correlation than VIC-EB. This is interesting, as VIC-WB is programmed to use the same energy balance as VIC-EB for the snow simulations and should predict similar values. The median of only the positive scores is more similar between both VIC models than the median of all the scores (Table 5), which means that the worse performing cells perform different in both models.

Assuming that the ASMR-E observations have a perfect match with reality, the argument could be made for both models that they are the better snow predictors. For the whole world PCR-GLOBWB is (slightly) more accurate than both VIC models, as the boxplots show that the median of the KGE and NS scores is (slightly) higher, and this model has fewer negative extreme values. On the other hand, VIC scores higher for all regions with positive scores and would therefore be the better choice for these regions. A more accurate control dataset, or better checks of the ASMR-E dataset could remove uncertainty about the actual SWE and lead to a better comparison between VIC and PCR-GLOBWB.

The last regional component that was compared in this study was the soil moisture. In contrast to the SWE, for soil moisture there are differences between PCR-GLOBWB and VIC. However, it is unlikely that the energy balance is the main cause of these differences. Both VIC models score better than PCR-GLOBWB for dry and wet areas, but especially the fact that VIC-WB scores slightly better than VIC-EB indicates that the energy balance is not the main cause of this. It is more likely that the calibration of VIC has a positive influence on the soil moisture in VIC, leading to the higher efficiency scores for soil moisture. Figure 16 is a confirmation of this, as the two VIC models score better than PCR-GLOBWB for the Amazon (which is calibrated in VIC) and similar for the Congo (which is

uncalibrated). Moreover, both VIC models use the same soil parameter file. Therefore, it is understandable that the soil moisture results are very similar.

To conclude, PCR-GLOBWB compares better to both Fluxnet and ERA-Interim for the ET, VIC scores better for the soil moisture and the results of the SWE simulations are similar to each other. So despite the fact that the overall results of the local comparison are balanced between PCR-GLOBWB and VIC, the discharge results are in favor of VIC. An important factor is the calibration of VIC, as VIC scores higher than PCR-GLOBWB for each of the calibrated basins. Given the fact that the method of calibration was aimed at recreating the runoff fields with VIC it is not surprising that VIC obtains higher scores for the discharge. However, the uncalibrated PCR-GLOBWB scores high for the Brahmaputra, Mackenzie and Magdalena (with the Penman-Monteith PET method also for the Amazon, see Appendix 1) and it would be interesting to see the scores of PCR-GLOBWB when this model is also calibrated.

Given the differences between PCR-GLOBWB and VIC in model structure, it is impossible to draw decisive conclusions about the effect of an energy balance when only these two models are compared. The VIC-EB and VIC-WB comparison is more suited for this. VIC-EB has higher efficiency scores for the ET (for both the ERA-Interim and Fluxnet comparison) and similar scores for the soil moisture, SWE and the discharge of most of the studied rivers. Based on this it can be concluded that an energy balance has a positive effect on the ET results of the model, but a different ET method (e.g. the method of PCR-GLOBWB) is more accurate compared to the Fluxnet and ERA-Interim datasets.

4.2. In which climate zone will this effect be the largest: tropic, arctic, or temperate?

The results show that PCR-GLOBWB scores higher for the ET in both areas where the energy balance was thought to be an advantage: the tropics and the boreal regions. VIC-EB scores better for both regions than VIC-WB, but PCR-GLOBWB outperforms both models.

A downside to this comparison is the fact that the tropical region has low (negative on some places) correlation scores for all three models. With ERA-Interim the correlations increase, but they remain low. Especially for the Amazon the correlation is negative for a large part of the basin. This, coupled with the low KGE and NS scores, indicates that there is a large discrepancy between the output of VIC and PCR-GLOBWB and the ET of Fluxnet and ERA-Interim. Since ET is an important factor of the hydrological balance in the Amazon, this should also be seen in the discharge. However, the discharge graphs of the Amazon show that the VIC models are able to approach the observed discharge better than the PCR-GLOBWB models. This could indicate that the ET observed by Fluxnet and ERA-Interim is less accurate than the ET modelled by VIC or it confirms the fact that calibration has a large effect on the accuracy of the discharge. This could be tested by simulating the discharge of other tropical rivers, which are not calibrated, and by comparing the ET output of the models to other ET datasets.

The output of the models is very similar for the SWE. This indicates that the effect of an energy balance in the snow module is small. However, snow melt plays a major role in the peak discharge of arctic rivers and both VIC models give a very good approximation of the peak discharge of the Mackenzie. The peak in PCR-GLOBWB is lower, indicating that the model has trouble with the snow melt peak during the spring. The discharge in PCR-GLOBWB after the peak is higher than in VIC and eventually also higher than the GRDC, with similar yearly discharge values. This means that the

response in PCR-GLOBWB is too slow and the water is too much retarded. In VIC, the response is actually too fast, as the discharge after the peak is lower than the observations of the GRDC. The response of the model is related to the soil parameters, since these determine the timing of the runoff and baseflow to the rivers. When the water is in the rivers, the routing module determines the velocity in the river, but this effect is small on a monthly timescale. Since both VIC models use the energy balance in the case of snow, the discharge is very similar in the winter. During the summer this is not the case and small differences between the predicted discharges of VIC-EB and VIC-WB are visible. In this period VIC-EB is somewhat more accurate than VIC-WB.

In conclusion, the calibrated soil parameters of VIC have a large influence on the accuracy of the discharge in the tropical and arctic basins. When only VIC-EB and VIC-WB are considered, it becomes clear that the energy balance has a positive effect on the ET and discharge in the tropics. In the arctic regions this could not be evaluated, as both VIC-EB and VIC-WB use the energy balance in the case of snow. VIC-WB does not use the energy balance in the summer and in this period the discharge scores are higher for VIC-EB in the Mackenzie. The ET scores are also higher for VIC-EB, which is also caused by the energy balance. However, the differences are larger over the tropics, where the magnitude of the energy based processes (ET) is larger and therefore the importance to model these processes more accurately.

4.3. How do VIC and PCR-GLOBWB perform in areas where abundant other information is available? And which model provides better estimates of water fluxes in regions where other information is not readily available?

The areas with abundant information are assumed to be the developed nations for the local variables and for the discharge the basins of the three big rivers; the Amazon, Mississippi and the Mackenzie. The Fluxnet and ERA-Interim dataset are both similar in the fact that they use observations coupled with a model to create a global grid of ET values (see data section). However, there still remain areas in the tropics with low correlations between both datasets and VIC and PCR-GLOBWB, indicating possible errors in any of the models. Given the dependence on observations in both datasets, which are focused in the developed world, it can be assumed that the accuracy of both ET datasets is higher for these regions. This might also be the reason why the accuracy scores are lower for the tropical regions.

The soil moisture and SWE data are provided by satellites. This means that there is no extra information or accuracy for the developed world compared to the developing world. It is possible that the extra information content of the information rich areas is included in the model side of the comparison, by using more accurate parameters in those areas. However, based on the results of both comparisons, there appears to be no clear bias towards the information rich areas. Therefore it can be concluded that both models perform similar for the whole world.

The discharge results do not show a division in EartH2Observe case study areas and information-rich areas. There is always at least one model that gives acceptable results for a single river and the highest scores occur in a river from the EartH2Observe program, the Brahmaputra. The worst performing rivers are the Mississippi and the Nile, where only VIC-WB obtains a positive Nash-Sutcliffe score. As shown in the results, the Nile has also very low scores for the local hydrological comparison; this explains the low scores for this basin.

A possible explanation why the results do not show a spatial distinction between information-rich and information-poor regions in the world is the use of globally available data as input. The most important parameters that describe the soil, vegetation, land use, climate etc. are globally available at the resolution of both models. Comparing the results of the models to accurate data with a higher resolution could result in more insight in these regional differences.

4.4. Will the results of VIC-WB be similar to those of PCR-GLOBWB, or are there other differences in the model structure of VIC that cause differences between the results of VIC-WB and PCR-GLOBWB?

The results have shown that, apart from the energy balance, other differences exist between VIC and PCR-GLOBWB that influence the results of the models. It can be seen from almost all results that VIC-WB corresponds more to VIC-EB than to PCR-GLOBWB. This indicates that the energy balance is not the only difference between both models. Other differences between both models that were identified before running the simulations were the ET calculations, the snow module, the calibration of VIC and the amount of subgrid variability in each cell. The effect of the amount of subgrid variability is difficult to measure and calibration has already been discussed above, but the other two large differences in model structure can be evaluated.

In order to compare the ET calculations to each other, a simulation was done where PCR-GLOBWB was forced with the potential ET calculated from the Penman-Monteith equation (see Appendix 1). However, despite the fact that VIC also calculates the potential ET with the Penman-Monteith equation, the results were still far apart. This indicates that there are other factors that also have influence on the ET. Given the fact that the ET depends on the amount of soil moisture and that these results differ between PCR-GLOBWB and VIC, it will not be possible to obtain exactly the same results for the ET. The gap between the results of VIC and PCR-GLOBWB shows that the fundamental equations to determine the actual ET in both models differ, leading to differences in the results of VIC-WB and PCR-GLOBWB.

Because VIC uses a different method to simulate snow melt and accumulation processes, it was not expected that the SWE results of VIC-WB were more comparable to PCR-GLOBWB than to VIC-EB. Since VIC-WB uses the same energy balance as VIC-EB for the snow simulations, it is not surprising that the results of VIC-WB are more similar to the results of VIC-EB. The discharge results confirm these findings, as the discharge of VIC-WB is more comparable to the discharge of VIC-EB for all basins. Only for the Mississippi are the results of VIC-WB closer to the results of PCR-GLOBWB than to the results of VIC-EB.

To sum up, the results of VIC-WB are more comparable to VIC-EB. The energy balance is not the only fundamental difference between both models and not the only reason why there are differences in results between VIC-EB and PCR-GLOBWB. Other reasons are the fundamental ET equations, the different snow module and the calibration/routing method used by the models.

5. Conclusion

This study evaluated the effect of an energy balance on the accuracy of a global hydrological model, as well as the sensitivity of PCR-GLOBWB to changes in climate forcing, PET forcing and resolution. This was done by 4 different simulations of PCR-GLOBWB, each with a different setting/forcing that was changed, and the two versions of VIC (EB and WB). The results were compared to global datasets of the main components of the hydrological balance: the evapotranspiration, the soil moisture, the snow water equivalent and the discharge. Based on this comparison, the simulations received local KGE, NS and correlation scores for the ET, SWE and soil moisture and regional scores for the discharge at six different measurement locations.

The main research question was to look at the effect of an energy balance in a global hydrological model. Compared to VIC-WB, the usage of an energy balance (VIC-EB) leads to higher accuracy scores for the ET and for three of the six rivers. The simulations with only the water balance score slightly higher for the soil moisture, but this difference is small and not in proportion to the increase of accuracy that the energy balance gives for the other components. Due to the parametrization of VIC, the results for the SWE are similar between VIC-EB and VIC-WB. PCR-GLOBWB performs better than both VIC models for the ET and only slightly lower for the soil moisture and SWE. The fact that PCR-GLOBWB scores similar for the SWE is surprising, especially given the fact that PCR-GLOBWB uses a relatively simple degree day snow module, while VIC has a more sophisticated melt/accumulation module with an energy balance. The very similar scores and patterns for the SWE suggest that the ASMR-E SWE dataset might not be accurate enough for a good comparison. The local results of the VIC comparison indicate that an energy balance has benefits for the ET and the SWE components of the water balance, but a completely different scheme (PCR-GLOBWB) might have even better or similar results. The results of PCR-GLOBWB can improve further if this model would introduce an energy balance. On the other hand, VIC can also benefit from the inclusion of the ET methods of PCR-GLOBWB.

Compared to PCR-GLOBWB, VIC performs better for the discharge, as it has the highest scores for five of the six rivers, VIC-EB for the Amazon, Brahmaputra and Mackenzie and VIC-WB for the Mississippi and the Nile. Only the Magdalena is simulated better by PCR-GLOBWB than by VIC. While the local results are comparable for both models, the discharge results are clearly in favour of VIC. Given the fact that both VIC models perform better than PCR-GLOBWB in this area, the difference is probably not due to the energy balance of VIC-EB. The main reason that VIC has a higher accuracy for the discharge is because five of the used basins were calibrated, the only basin that was not calibrated, the Magdalena, was better predicted in PCR-GLOBWB. Moreover, VIC-WB performs similar or better than VIC-EB for the discharge predictions of all basins, reducing the need for an energy balance if the focus of study is on discharge predictions.

The sensitivity of PCR-GLOBWB is not the same for each of the tested input/configuration. Forcing the model with the PET from the Penman-Monteith equation instead of the Hamon equation has a positive effect on all tested parts of the water balance. The discharge results are more similar to the results of the GRDC and to VIC, which also uses the Penman-Monteith equation for the PET. The effect of a different climate forcing can generally be found on a regional level. On a local level the changes are small, but the accumulated effect of the larger amount of precipitation in the GPCC dataset can be seen in the discharge of the studied rivers. All rivers show an increase in the

discharge, which leads to higher scores for the Amazon, Brahmaputra, Mackenzie and Magdalena. The Mississippi and the Nile are better predicted by the CRU precipitation.

The final comparison was done to see the effect of an increase of resolution. It was expected that by increasing the resolution of PCR-GLOBWB from 30 arcminutes to 5 arcminutes, the accuracy of the model would improve as well. The higher resolution led to worse comparison scores on a local level, but some improvement for the discharge of the Amazon, Brahmaputra, Mackenzie and Magdalena. The reason for this remains unclear, but might be related to the input or comparison data. A part of the data was down/upscaled to 5 arcminute resolution and this conversion might have caused some discrepancies in the data. However, the decline in accuracy is so large in some areas that it is more likely that there exists an error in the model.

Further research

Areas where further research could provide more definite answers related to the effect of the energy balance are the ASMR-E SWE dataset and the effect of the Penman-Monteith PET on other areas in the world. As previously stated, the results of the SWE simulated by the models and the SWE found by ASMR-E differ on several places in the world. Given the strong correlation between the models and ASMR-E and between VIC and PCR-GLOBWB themselves, two models with fundamentally different snow modules, there is doubt about the accuracy of ASMR-E in the areas where the KGE and NS scores are low for both models. A different SWE dataset could be used to see if this would change any of the outcomes of this study. For now, VIC performs better in the areas where both models score high and worse for the areas where they do not. It would be a strong point in favor of the energy balance of VIC if it would turn out to be more accurate in the entire snow covered area. The use of Penman-Monteith for a global simulation of PCR-GLOBWB would result in a better worldwide comparison between VIC and PCR-GLOBWB by eliminating a component that is different in both models.

6. References

Adler, R. F., Kidd, C., Petty, G., Morissey, M., & Goodman, H. M. (2001). Intercomparison of global precipitation products: The third Precipitation Intercomparison Project (PIP-3). Bulletin of the American Meteorological Society, 82(7), 1377-1396.

Alcamo, J., Flörke, M., & Märker, M. (2007). Future long-term changes in global water resources driven by socio-economic and climatic changes. Hydrological Sciences Journal, 52(2), 247-275.

Beltaos, S. (2008). Progress in the study and management of river ice jams. Cold Regions Science and Technology, 51(1), 2-19.

Bergström, S., & Singh, V. P. (1995). The HBV model. Computer models of watershed hydrology., 443-476.

Beven, K. J., & Cloke, H. L. (2012). Comment on "Hyperresolution global land surface modeling: Meeting a grand challenge for monitoring Earth's terrestrial water" by Eric F. Wood et al. Water Resources Research, 48(1).

Biemans, H., Hutjes, R. W. A., Kabat, P., Strengers, B. J., Gerten, D., & Rost, S. (2008). Effects of precipitation uncertainty on discharge calculations for main river basins. Journal of Hydrometeorology, 10(4), 1011-1025.

Bowling, L. C., & Lettenmaier, D. P. (2010). Modeling the effects of lakes and wetlands on the water balance of Arctic environments. Journal of Hydrometeorology, 11(2), 276-295.

Davie, J. C. S., Falloon, P. D., Kahana, R., Dankers, R., Betts, R., Portmann, F. T., ... & Arnell, N. (2013). Comparing projections of future changes in runoff from hydrological and biome models in ISI-MIP. Earth System Dynamics, 4, 359-374.

Dee, D. P., Uppala, S. M., Simmons, A. J., Berrisford, P., Poli, P., Kobayashi, S., ... & Bechtold, P. (2011). The ERA-Interim reanalysis: Configuration and performance of the data assimilation system. Quarterly Journal of the Royal Meteorological Society, 137(656), 553-597.

Dorigo, W. A., Gruber, A., De Jeu, R. A. M., Wagner, W., Stacke, T., Loew, A., ... & Kidd, R. (2015). Evaluation of the ESA CCI soil moisture product using ground-based observations. Remote Sensing of Environment, 162, 380-395.

Dutra, E., Balsamo, G., Calvet, J., Minvielle, M., Eisner, S., Fink, G., Pessenteiner, S., Orth, R., Burke, S., Van Dijk, A., Polcher, J., Beck, H., de La Torre, A. M. (2015). Report on the current state-of-the-art Water Resources Reanalysis.

Emmerton C.A., Lesack L.F.W., Marsh P. 2007. Lake abundance, potential water storage, and habitat distribution in the Mackenzie River Delta, Western Canadian Arctic. Water Resources Research 43: 14.

Fekete, B. M., Vörösmarty, C. J., Roads, J. O., & Willmott, C. J. (2004). Uncertainties in precipitation and their impacts on runoff estimates. Journal of Climate, 17(2), 294-304.

Food and Agriculture Organization of the United Nations (FAO), (2003) Digital Soil Map of the World, Version 3.6., FAO, Rome, Italy, available at: http://data.fao.org/map?entryId=446ed430-8383-11db-b9b2-000d939bc5d8 (on 14/07/15)

Franchini, M., and M. Pacciani (1991), Comparative-analysis of several conceptual rainfall runoff models, Journal of Hydrology, 122(1-4), 161-219.

Gain, A. K., Immerzeel, W. W., Sperna Weiland, F. C., & Bierkens, M. F. P. (2011). Impact of climate change on the stream flow of the lower Brahmaputra: trends in high and low flows based on discharge-weighted ensemble modelling. Hydrology and Earth System Sciences, 15(5), 1537-1545.

Gao, H., Q. Tang, X. Shi, C. Zhu, T. J. Bohn, F. Su, J. Sheffield, M. Pan, D. P. Lettenmaier, and E. F. Wood, 2010a: Water Budget Record from Variable Infiltration Capacity (VIC) Model. In Algorithm Theoretical Basis Document for Terrestrial Water Cycle Data Records (in review)

Gao, Y., Xie, H., Lu, N., Yao, T., & Liang, T. (2010b). Toward advanced daily cloud-free snow cover and snow water equivalent products from Terra–Aqua MODIS and Aqua AMSR-E measurements. Journal of Hydrology, 385(1), 23-35.

Goolsby, D.A. (2000). Mississippi basin nitrogen flux believed to cause Gulf hypoxia. Eos, Transactions American Geophysical Union, 81(29), 321-327.

Gosling, S., Taylor, R. G., Arnell, N., & Todd, M. C. (2011). A comparative analysis of projected impacts of climate change on river runoff from global and catchment-scale hydrological models. Hydrology and Earth System Sciences, 15(1), 279-294.
Goulding, H. L., Prowse, T. D., & Beltaos, S. (2009). Spatial and temporal patterns of break-up and ice-jam flooding in the Mackenzie Delta, NWT. Hydrological processes, 23(18), 2654-2670.
Gottschalk, L., Krasovskaia, I., Dominguez, E., Caicedo, F., & Velasco, A. (2015). Interpolation of monthly runoff along rivers applying empirical orthogonal functions: application to the Upper Magdalena River, Colombia. Journal of Hydrology.
Gudmundsson, L., Tallaksen, L. M., Stahl, K., Clark, D. B., Dumont, E., Hagemann, S., ... & Koirala, S. (2011). Comparing large-scale hydrological model simulations to observed runoff percentiles in Europe. Journal of Hydrometeorology, 13(2), 604-620.
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), 80-91.
Haddeland, I., Skaugen, T., & Lettenmaier, D. P. (2006a). Anthropogenic impacts on continental surface water fluxes. Geophysical Research Letters, 33(8).
Haddeland, I., Lettenmaier, D. P., & Skaugen, T. (2006b). Effects of irrigation on the water and energy balances of the Colorado and Mekong river basins. Journal of Hydrology, 324(1), 210-223.
Haddeland, I., Clark, D. B., Franssen, W., Ludwig, F., VOß, F. R. A. N. K., Arnell, N. W., ... & Yeh, P. (2011). Multimodel estimate of the global terrestrial water balance: Setup and first results. Journal of Hydrometeorology, 12(5), 869-884.
Hagemann, S. and Gates, L. D. (2003). Improving a subgrid runoff parameterization scheme for climate models by the use of high resolution data derived from satellite observations. Climate Dynamics, 21(3-4), 349-359.
Harris, I. P. D. J., Jones, P. D., Osborn, T. J., & Lister, D. H. (2014). Updated high-resolution grids of monthly climatic observations–the CRU TS3. 10 Dataset. International Journal of Climatology, 34(3), 623-642.
Hurkmans, R. T. W. L., De Moel, H., Aerts, J. C. J. H., & Troch, P. A. (2008). Water balance versus land surface model in the simulation of Rhine river discharges. Water Resources Research, 44(1).
Jung, M., Reichstein, M., & Bondeau, A. (2009). Towards global empirical upscaling of FLUXNET eddy covariance observations: validation of a model tree ensemble approach using a biosphere model. Biogeosciences, 6(10), 2001-2013.
Kraijenhoff Van de Leur, D. A. (1958). A study of non-steady groundwater flow with special reference to a reservoir coefficient. De Ingenieur, 70(19), 87-94.
Krause, P., Boyle, D. P., & Bäse, F. (2005). Comparison of different efficiency criteria for hydrological model assessment. Advances in Geosciences, 5, 89-97.
Liang, X., Lettenmaier, D. P., Wood, E. F., & Burges, S. J. (1994). A simple hydrologically based model of land surface water and energy fluxes for general circulation models. JOURNAL OF GEOPHYSICAL RESEARCH-ALL SERIES-, 99, 14-415.
Liu, Y. Y., Parinussa, R. M., Dorigo, W. A., De Jeu, R. A. M., Wagner, W., van Dijk, A. I. J. M., McCabe, M. F., Evans, J. P. (2011). Developing an improved soil moisture dataset by blending passive and active microwave satellite-based retrievals. Hydrology and Earth System Sciences, 15, 425-436
Liu, Y. Y., Dorigo, W. A., Parinussa, R. M., De Jeu, R. A. M., Wagner, W., McCabe, M. F., ... & Van Dijk, A. I. J. M. (2012). Trend-preserving blending of passive and active microwave soil moisture retrievals. Remote Sensing of Environment, 123, 280-297.
Lohmann, D., R. Nolte-Holube, and E. Raschke, 1996: A large-scale horizontal routing model to be coupled to land surface parametrization schemes, Tellus, 48(A), 708-721.
Lohmann, D., E. Raschke, B. Nijssen and D. P. Lettenmaier, 1998: Regional scale hydrology: I. Formulation of the VIC-2L model coupled to a routing model, Hydrol. Sci. J., 43(1), 131-141.
Martinez, J. M., Guyot, J. L., Filizola, N., & Sondag, F. (2009). Increase in suspended sediment discharge of the Amazon River assessed by monitoring network and satellite data. Catena, 79(3),

257-264.
McCuen, R. H., & Snyder, W. M. (1975). A proposed index for comparing hydrographs. Water Resources Research, 11(6), 1021-1024.
Mueller, B., Seneviratne, S. I., Jimenez, C., Corti, T., Hirschi, M., Balsamo, G., ... & Jung, M. (2011). Evaluation of global observations-based evapotranspiration datasets and IPCC AR4 simulations. Geophysical Research Letters, 38(6).
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.
Murphy, A. H. (1988). Skill scores based on the mean square error and their relationships to the correlation coefficient. Monthly weather review, 116(12), 2417-2424.
Nash, J. E. and Sutcliffe, J. V.: River flow forecasting through conceptual models, Part I - A discussion of principles, J. Hydrol., 10, 282–290, 1970.
Oak Ridge National Laboratory Distributed Active Archive Center (ORNL DAAC). 2015. FLUXNET Web Page. Available online [http://fluxnet.ornl.gov] from ORNL DAAC, Oak Ridge, Tennessee, U.S.A. Accessed March 4, 2016
Penn, J. R. (2001). Rivers of the world: a social, geographical, and environmental sourcebook. Abc-clio.
Prudhomme, C., Giuntoli, I., Robinson, E. L., Clark, D. B., Arnell, N. W., Dankers, R., ... & Wisser, D. (2014). Hydrological droughts in the 21st century, hotspots and uncertainties from a global multi-model ensemble experiment. Proceedings of the National Academy of Sciences, 111(9), 3262-3267.
Schneider, U., Becker, A., Finger, P., Meyer-Christoffer, A., Ziese, M., & Rudolf, B. (2014). GPCC's new land surface precipitation climatology based on quality-controlled in situ data and its role in quantifying the global water cycle. Theoretical and Applied Climatology, 115(1-2), 15-40.
Sood, A., & Smakhtin, V. (2015). Global hydrological models: a review. Hydrological Sciences Journal, 60(4), 549-565.
Storck, P., et al. (2002), Measurement of snow interception and canopy effects on snow accumulation and melt in a mountainous maritime climate, Oregon, United States, Water Resour Res, 38(11), 16.
Tedesco, M., R. Kelly, J. L. Foster, and A. T. Chang. 2004. AMSR-E/Aqua Daily L3 Global Snow Water Equivalent EASE-Grids, Version 2. Monthly dataset. Boulder, Colorado USA. NASA National Snow and Ice Data Center Distributed Active Archive Center. doi: http://dx.doi.org/10.5067/AMSR-E/AE_DYSNO.002. [11/11/2015].
Van Beek, L.P.H. and M.F.P. Bierkens (2008), The Global Hydrological Model PCR-GLOBWB: Conceptualization, Parameterization and Verification, Report Department of Physical Geography, Utrecht University, Utrecht, The Netherlands, http://vanbeek.geo.uu.nl/suppinfo/vanbeekbierkens2009.pdf.
Van Beek, L. P. H., Wada, Y., & Bierkens, M. F. (2011). Global monthly water stress: 1. Water balance and water availability. Water Resources Research, 47(7).
Van Beek, L. P. H., Eikelboom, T., Vliet, M. T., & Bierkens, M. F. (2012). A physically based model of global freshwater surface temperature. Water Resources Research, 48(9).
Wada, Y., Van Beek, L. P. H., Wanders, N., & Bierkens, M. F. (2013). Human water consumption intensifies hydrological drought worldwide. Environmental Research Letters, 8(3), 034036.
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 Discussions, 5(1), 15-40.
Wagner, W., W. Dorigo, R. de Jeu, D. Fernandez, J. Benveniste, E. Haas, M. Ertl (2012). Fusion of active and passive microwave observations to create an Essential Climate Variable data record on soil moisture. ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information

Sciences (ISPRS Annals), Volume I-7, XXII ISPRS Congress, Melbourne, Australia, 25 August-1 September 2012, 315-321
Weedon, G. P., G. Balsamo, N. Bellouin, S. Gomes, M. J. Best, and P. Viterbo (2014), The WFDEI meteorological forcing data set: WATCH Forcing Data methodology applied to ERA-Interim reanalysis data, Water Resour. Res., 50, 7505–7514, doi:10.1002/2014WR015638.
Wood, E. F., Roundy, J. K., Troy, T. J., Van Beek, L. P. H., Bierkens, M. F., Blyth, E., ... & Gochis, D. (2011). Hyperresolution global land surface modeling: Meeting a grand challenge for monitoring Earth's terrestrial water. Water Resources Research, 47(5).
Wood, E. F., Roundy, J. K., Troy, T. J., Beek, R., Bierkens, M., Blyth, E., ... & Gochis, D. (2012). Reply to comment by Keith J. Beven and Hannah L. Cloke on "Hyperresolution global land surface modeling: Meeting a grand challenge for monitoring Earth's terrestrial water". Water Resources Research, 48(1).
Zhou, T., Nijssen, B., Gao, H., & Lettenmaier, D. P. (2016). The Contribution of Reservoirs to Global Land Surface Water Storage Variations*. Journal of Hydrometeorology, 17(1), 309-325.

A. Appendix 1: Sensitivity of PCR-GLOBWB to changes in input forcing, resolution and evapotranspiration method

Here, the results will be described and discussed that are not directly related to the VIC-PCR-GLOBWB model structure comparison. These comparisons were related to the final research question: How sensitive is PCR-GLOBWB to changes in climate forcing, potential ET forcing and resolution?

Climate input

In order to see where the differences between the CRU and the GPCC precipitation datasets are located, a small evaluation of these datasets was done. Both datasets were compared for the period 1979-2010 on a daily basis. Figure A1a shows that, on average, the daily precipitation differences are not very large and remain below 1 mm/day for the majority of the world. However, there are some local differences where the discrepancy between both datasets is larger. Moreover, the global sum of the average daily difference between GPCC and CRU is 1099mm in favor of the GPCC. Meaning that using the GPCC dataset leads to 1099mm water extra each day around the world. This includes the area over Greenland where CRU precipitation values are higher. Especially for the Amazon, the amount of water available with GPCC is larger than with CRU. Müller Schmied et al. (2014) also did some research about the differences between CRU TS 3.2 and GPCC v6 precipitation and their study found that the use of the GPCC dataset leads to an average increase of 8.7% compared to the CRU precipitation for each cell. Moreover, on 37.5% of the land area (except Greenland and Antarctica) the increase is more than 10%.

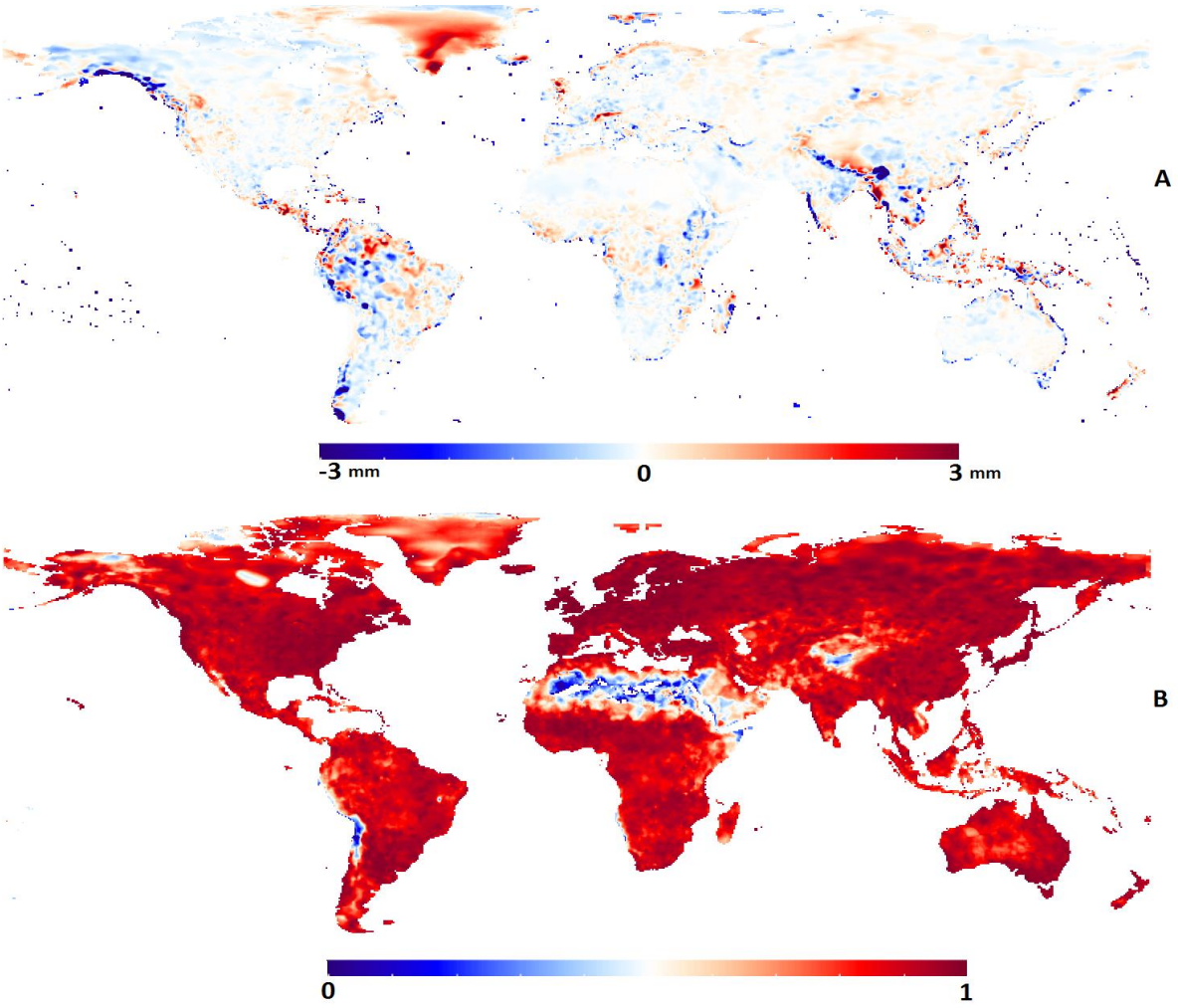


Figure A1: A: Average daily difference between the CRU and GPCC precipitation in mm/day for each grid cell. A positive value indicates that the CRU has on average more precipitation than the GPCC in that cell and vice versa. B: Correlation between CRU and GPCC. All cells have a positive correlation.

Other regions of interest are the areas where the correlation between the precipitation datasets is relatively low (Figure A1B). The correlation is positive for the whole world, but there are several areas where the correlation between the datasets is below 0.5. These areas are the Andes, the Himalayas/Tibetan plateau, Arabia/Sahara and two areas in Canada/Alaska. Some of the areas where the absolute difference in precipitation is relatively large show a strong correlation.

Table A1: Median KGE, NSE and correlation scores for the precipitation input simulations. Simulations were done with PCR-GLOBWB on 30 arcminute resolution with different precipitation datasets (CRU and GPCC).

		Kling-Gupta efficiency		Nash-Sutcliffe efficiency		Correlation	
		CRU	GPCC	CRU	GPCC	CRU	GPCC
ET	Fluxnet	0.47	0.46	0.45	0.41	0.92	0.89
	ERA-Interim	0.43	0.44	0.31	0.32	0.87	0.90
SM	ESA CCI	0.21	0.21	x	x	0.59	0.59
SWE	ASMR-E	-0.05	-0.03	-0.13	-0.09	0.70	0.69

Table A1 indicates that the difference between the two precipitation datasets is small. The scores are very close to each other and when the output is directly compared the KGE, NSE and the correlations are high (Table A2). The scores indicate that the output generated by the GPCC forcing data is more comparable to the ET-ERA and ASMR-E SWE data, whereas the CRU output corresponds better to the ET-FLUX. Even though the differences between the scores are very small, the differences for the discharge are larger. Figure A3 clearly shows that the six rivers have a higher discharge in the GPCC simulations. There are some deviations in the discharge patterns, but these changes remain small. For some rivers this leads to better scores (Amazon, Brahmaputra, Mackenzie and Magdalena) and for others to worse scores (Mississippi and the Nile) (Table A3).

The spatial plots (Figure A2) show the areas with the largest differences between the results of the two PCR-GLOBWB simulations. The main areas with low KGE and correlation scores are located around the Sahara, Arabia, Central-Asia and Canada/Alaska, which are the same areas where the differences between the precipitation datasets were relatively large. The spatial plots also allow for evaluation of the river basins, in order to see what causes the change in discharge patterns. For the ET, almost all basins have high KGE and correlation scores, apart from the Amazon. The northern part of the Nile basin has low KGE scores as well, but the values next to the river are higher and play a larger role in the hydrology of the river. For the whole world, the SM scores are lower, especially in the Mackenzie basin.

Table A2: Median KGE, NSE and correlation scores when the output of PCR-GLOBWB with the CRU precipitation dataset is compared to the output of PCR-GLOBWB with the GPCC dataset.

	Kling-Gupta efficiency	Nash-Sutcliffe efficiency	Correlation
ET	0.90	0.87	0.94
SM	0.74	0.53	0.81
SWE	0.63	0.77	0.96

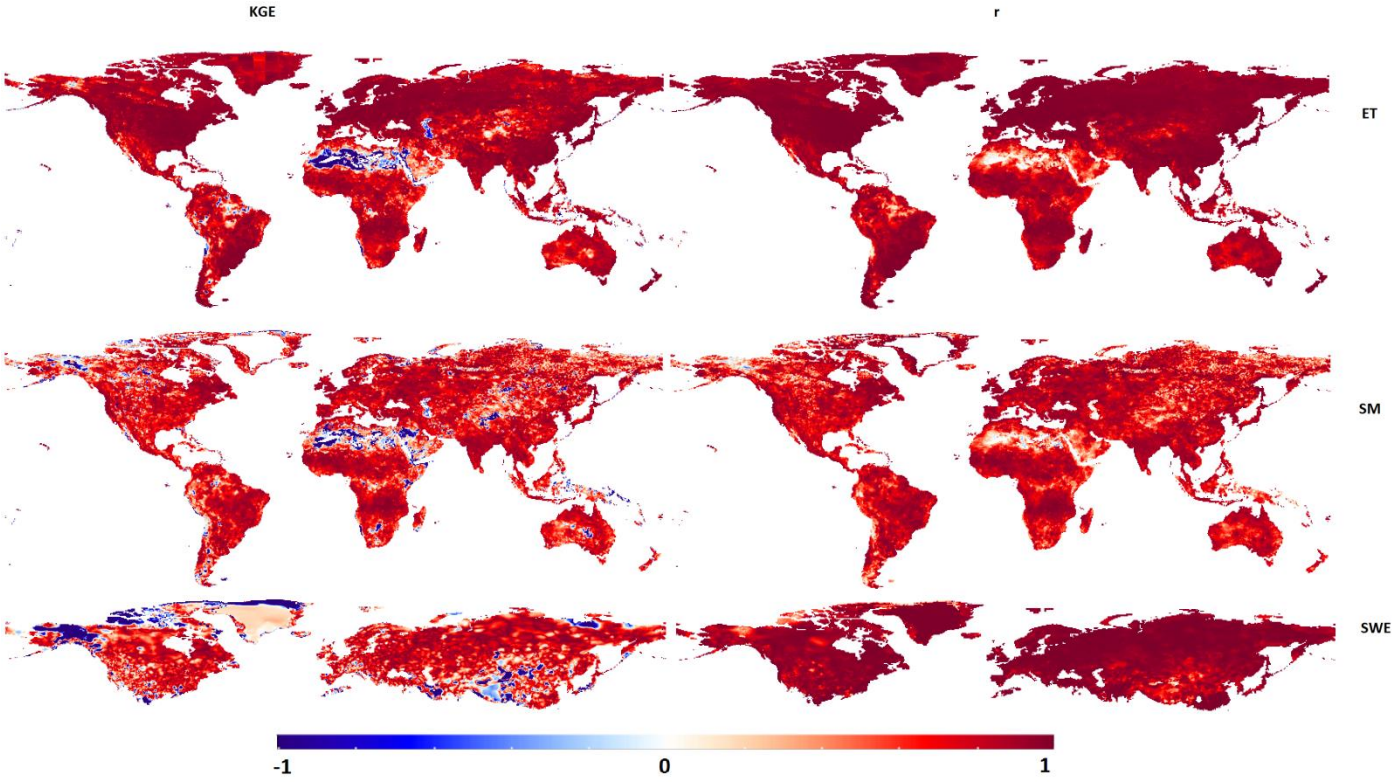


Figure A2: KGE (left) and correlation (right) scores when comparing the GPCC PCR output with the CRU PCR output for all the three comparison criteria (top to bottom: ET, SM and SWE).

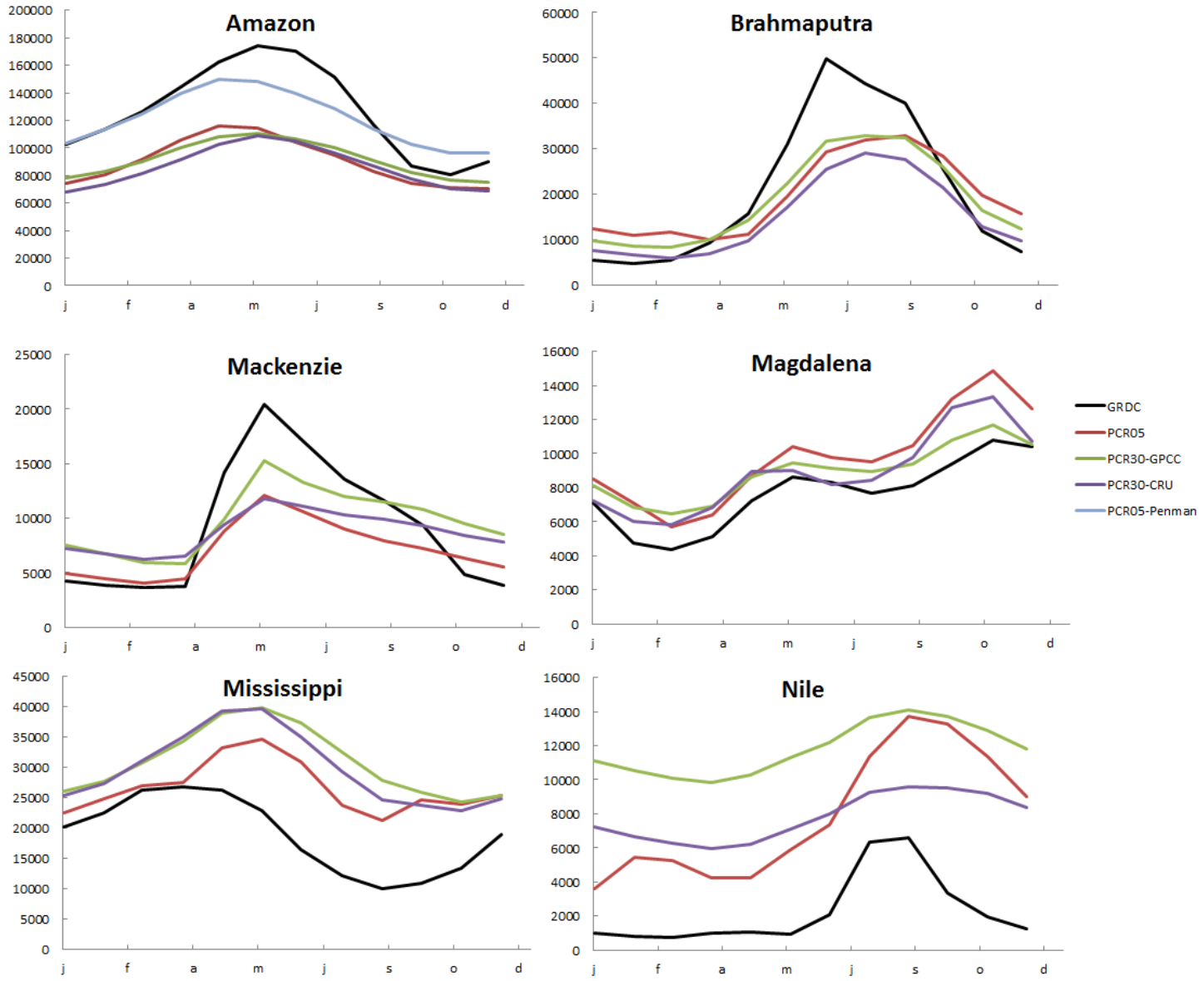


Figure A3: Discharge (in m^3/s) seasonality for the selected rivers for the period for which data is available from the GRDC. These periods are: for the Amazon (at Jatuares): 1992-2010, for The Brahmaputra (at Bahadurabad): 1985-2010, for the Mackenzie (at Arctic Red River): 1982-2010, for the Magdalena (at Calamar): 1980-1990, for the Mississippi (at Vicksburg): 1980-2010 and for the Nile (at Dongola): 1982-2002. PCR05 are the PCR-GLOBWB simulations with a resolution of 5 arcminutes and CRU precipitation, PCR30-GPCC are the PCR-GLOBWB simulations with a resolution of 30 arcminutes and GPCC precipitation, PCR30-CRU are the PCR-GLOBWB simulations with a resolution of 30 arcminutes with the CRU precipitation and PCR05-Penman are the PCR-GLOBWB simulations with a resolution of 5 arcminutes with the CRU precipitation and the Penman-Monteith PET.

Table A3: Kling-Gupta, Nash-Sutcliffe efficiency scores and the coefficient of determination of the discharge at an observation point in the simulated river basins for the two simulations of PCR-GLOBWB. PCR-CRU corresponds to the PCR-GLOBWB simulation with the CRU precipitation as input and PCR-GPCC to the PCR-GLOBWB simulation with the GPCC precipitation as input.

		Kling-Gupta efficiency		Nash-Sutcliffe efficiency		r^2	
Basin *(location)*	Model (mode)	PCR-CRU	PCR-GPCC	PCR-CRU	PCR-GPCC	PCR-CRU	PCR-GPCC
Amazon *(Jatuares)*		0.38	0.36	-1.80	-0.50	0.38	0.80
Brahmaputra *(Bahadurabad)*		0.43	0.56	0.55	0.74	0.83	0.88
Mackenzie *(Arctic Red River)*		0.35	0.53	0.46	0.62	0.67	0.71
Magdalena *(Calamar)*		0.72	0.69	0.41	0.57	0.70	0.89
Mississippi *(Vicksburg)*		0.35	0.35	-1.10	-1.58	0.47	0.36
Nile *(Dongola)*		-1.59	-3.23	-7.00	<-10	0.26	0.61

Potential evapotranspiration (PET)

The Amazon basin was used as the test area for the PET. PCR-GLOBWB was run a second time on 5 arcminute resolution, but with the PET forced by an input dataset. The reference potential ET calculated by the Penman-Monteith (hereafter Penman) was used in comparison with the Hamon PET calculated by a previous simulation. The CRU precipitation was used for both simulations.

Table A4: Median KGE, NSE and correlation scores for the Amazon basin with two different evapotranspiration methods.

	Kling-Gupta efficiency		Nash-Sutcliffe efficiency		Correlation	
	Hamon	Penman	Hamon	Penman	Hamon	Penman
ET Fluxnet	-0.42	-0.27	-22.5	-4.47	0.34	0.63
SM ESA CCI	0.10	0.17	-2.46	-2.28	0.63	0.59

The Amazon basin is one of the places with the lowest ET scores, compared to the rest of the world. With the Penman PET, the scores still remain below zero. Despite the fact that the model still fails the 'better than the mean of the observations' benchmark, there is improvement when PCR-GLOBWB is forced with the Penman-Monteith PET data (Figure A4 and Table A4). Especially in the eastern part of the basin the scores have increased considerably. The NS scores do not show a large improvement at first glance, but the boxplots (Figure A5) show that most NS scores are less negative with the Penman PET than with the Hamon ET. This is not shown in Figure A4, as all values below -1 are marked as -1 to keep visible differentiation of the most important scores (-1 to 1) possible. Another large difference is the correlation in the eastern Amazon. The Hamon PET has a negative correlation for large parts in this area, but the Penman PET shows a higher correlation for this area.

The third column of Figure A4 shows the KGE, NS and correlation scores when the outputs of PCR-GLOBWB of the two PET methods are compared to each other. It shows that there are structural differences between both datasets. Low KGE and NS scores over the Andes and the Amazon indicate that the output between the models is not comparable in those areas. High correlation scores occur in parts of these areas, but especially in the eastern part of the Amazon the correlation is below zero, implying that the trend for these cells is opposite to each other. See Table A5 for the median scores for this comparison.

With the more accurate ET, the discharge is also simulated more accurately by PCR-GLOBWB. Figure A3 and Table A6 show that the efficiency scores are higher in the Penman-Monteith PET simulation. The figure shows that the average discharge of the Amazon is higher in this simulation and the seasonality of the basin is more pronounced.

In Figure A5 the boxplots are shown with the scores of the PCR-GLOBWB output. Since VIC simulates the PET using the Penman-Monteith equation, the results of VIC are also included in this figure for comparison. Please note that VIC uses the GPCC precipitation instead of the CRU precipitation of PCR-GLOBWB and that VIC was originally run on 30 arcminute resolution. For this comparison the results of VIC were upscaled to 5 arcminutes and compared to the Fluxnet. Note that the Amazon was one of the regions where the precipitation differed between the GPCC and CRU datasets, so this is not a perfect comparison.

It is clear that, for the Amazon basin, PCR-GLOBWB scores more accurately compared to the Fluxnet when the PET is forced by the Penman-Monteith potential ET. VIC also uses the Penman-Monteith equation to calculate the potential ET and uses this to find the actual ET for the three components (canopy evaporation, vegetation transpiration and bare soil evaporation) (Gao et al., 2010a). However, the boxplots in Figure A5, but also Figure 12 and Figure 13, show that the VIC output corresponds less to the Fluxnet and ERA-Interim database than PCR-GLOBWB.

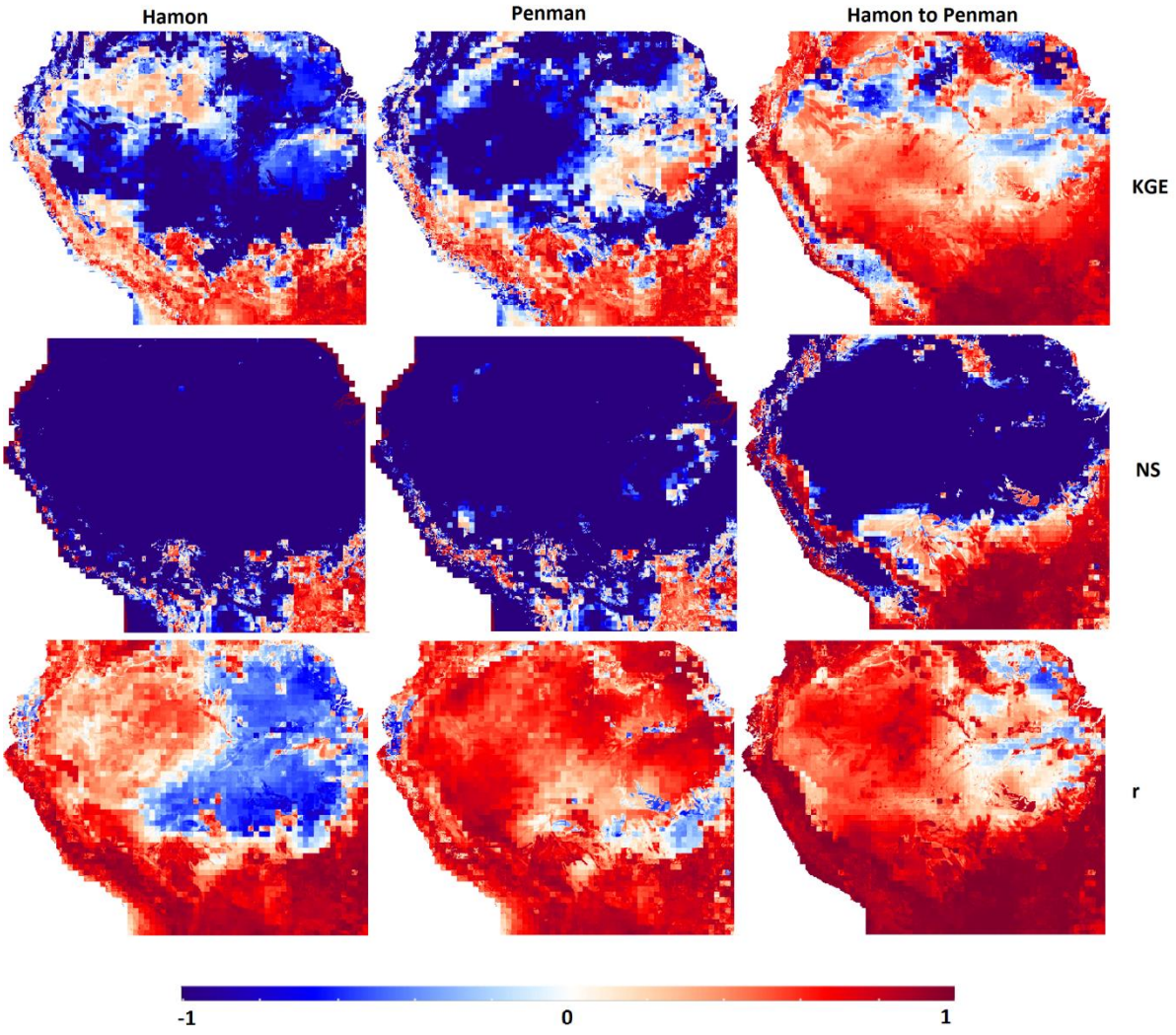


Figure A4: Spatial plots with the KGE, NS and correlation scores of the two different PCR-GLOBWB simulations. The third column shows the scores when the output of the Hamon run is directly compared to the output of the Penman-Monteith run.

Table A5: Median KGE, NSE and correlation scores when comparing the output of PCR-GLOBWB with the Hamon PET method to the Penman-Monteith PET method.

	Kling-Gupta efficiency	Nash-Sutcliffe efficiency	Correlation
ET Fluxnet	0.43	-1.96	0.67
SM	0.87	0.84	0.93

Table A6: Kling-Gupta and Nash-Sutcliffe efficiency scores and the coefficient of determination of the discharge at an observation point in the Amazon basin.

		Kling-Gupta efficiency		Nash-Sutcliffe efficiency		r^2	
Basin *(location)*	Model (mode)	Hamon	Penman-Monteith	Hamon	Penman-Monteith	Hamon	Penman-Monteith
Amazon *(Jatuares)*		0.45	0.64	-1.26	0.44	0.48	0.57

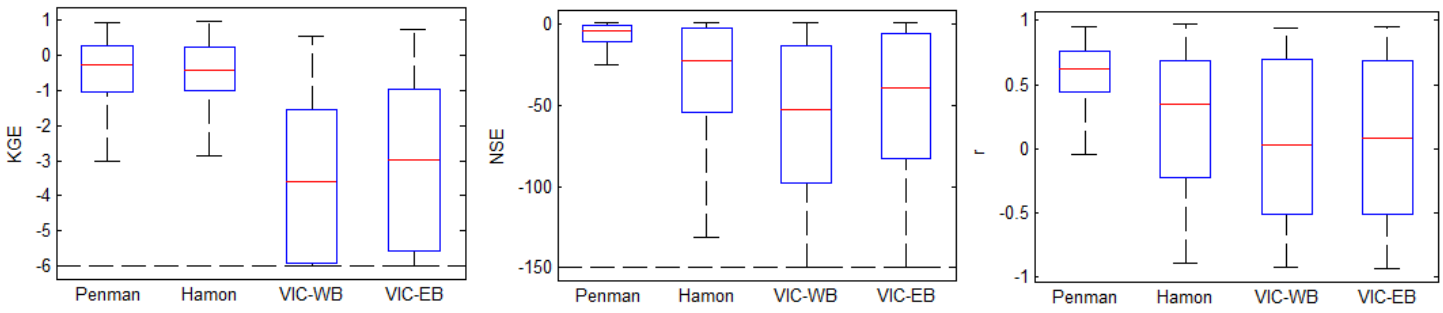


Figure A5: From left to right: KGE, NS and correlation scores for the runs of PCR-GLOBWB and VIC with Fluxnet as comparison dataset. Outliers are not shown in the boxplots. Both PCR-GLOBWB simulations (Penman and Hamon) were run with a resolution of 5 arcminutes and forced with the CRU precipitation dataset. Both VIC simulations were run with a resolution of 30 arcminutes and upscaled to 5 minutes before comparing it to the Fluxnet dataset. They were forced with the GPCC precipitation.

Resolution

Finally, to test the effect of a higher resolution, a comparison was made between PCR-GLOBWB on 30 arcminute resolution and on 5 arcminute resolution. The six case study areas were used for this, as a worldwide simulation would need too much computation time. This comparison used the CRU precipitation, therefore the PCR-30 scores correspond to the PCR-CRU scores reported in the Climate input sensitivity test earlier.

Table A7: Median KGE, NS and correlation scores for all studied basins in the two modes of PCR-GLOBWB. PCR-05 corresponds with the PCR-GLOBWB simulation on 5 arcminute resolution, while PCR-30 corresponds to the PCR-GLOBWB simulation on 30 arcminute resolution.

		Kling-Gupta efficiency		Nash-Sutcliffe efficiency		Correlation	
Basin		PCR-05	PCR-30	PCR-05	PCR-30	PCR-05	PCR-30
Amazon	ET-Flux	-0.42	-0.31	-22.5	-24.2	0.34	0.35
	SM	0.12	0.17	-2.31	-1.63	0.61	0.64
Brahmaputra	ET-Flux	0.28	0.51	-0.45	0.19	0.90	0.93
	SM	0.19	0.40	-4.29	-1.61	0.48	0.69
	SWE	-0.48	-0.50	-0.09	-0.07	0.24	0.22
Mackenzie	ET-Flux	0.71	0.65	0.85	0.81	0.97	0.97
	SM	0.01	0.06	x	x	0.26	0.27
	SWE	-0.04	-0.05	0.07	0.03	0.78	0.79
Magdalena	ET-Flux	-0.59	-0.27	-36.9	-40.2	0.49	0.45
	SM	0.31	0.40	-1.95	-1.25	0.56	0.63
Mississippi	ET-Flux	0.69	0.72	0.74	0.79	0.95	0.96
	SM	-0.09	0.012	-52.5	-19.02	0.29	0.39
	SWE	-0.10	-0.11	-0.07	-0.05	0.59	0.61
Nile	ET-Flux	0.16	0.21	-1.17	-1.00	0.68	0.71
	SM	-0.16	-0.05	-7.83	-4.70	0.43	0.46

It becomes clear from Table A7, Figure A6 and Figure A7 that, despite the higher resolution, the KGE, NS and correlation scores do not increase for almost all basins and test variables. They even decrease for the majority of the test cases. For the soil moisture this difference is generally larger than for the ET and SWE, but the pattern is similar. The Mackenzie is the only region that shows an increase in the ET accuracy when the resolution of the simulation is increased, all regions show a decrease in accuracy for the SM. However, despite the apparent decrease in accuracy with a higher resolution, Table A8 and Figure A3 show that several rivers obtain higher scores in the PCR-05 simulation (compared to PCR30-CRU). Apart from the Magdalena and the Nile, all rivers have higher KGE and NSE scores for the high resolution simulation.

The boxplots also show that not every basin has the same range of scores. The Amazon for example has correlation scores for ET between 1 and -1, whereas the Mackenzie has almost all scores between 1 and 0.8. The spread between the regions is generally the same for both resolutions. Only the Brahmaputra has a large difference between the two resolutions for the correlation. The median correlation does not decrease significantly (0.90 to 0.93), only the bottom 50% of the correlation scores is spread out more in the 5 resolution simulations. The spread of the scores for the different basins is more comparable to each other for the soil moisture.

For the Mackenzie basin more than half of the cells do not have enough values to obtain a NS score, resulting in a very low median. The reason for this is that these cells are frozen for several months, leading to NoData values in the satellite data and not enough unfrozen months remain to obtain a valid NS score. For the Brahmaputra and Mississippi less than 50% of the cell values are not useable, this leads to the long tail visible in Figure A7. The Kling-Gupta can be calculated for these regions, because of the differences in calculation method.

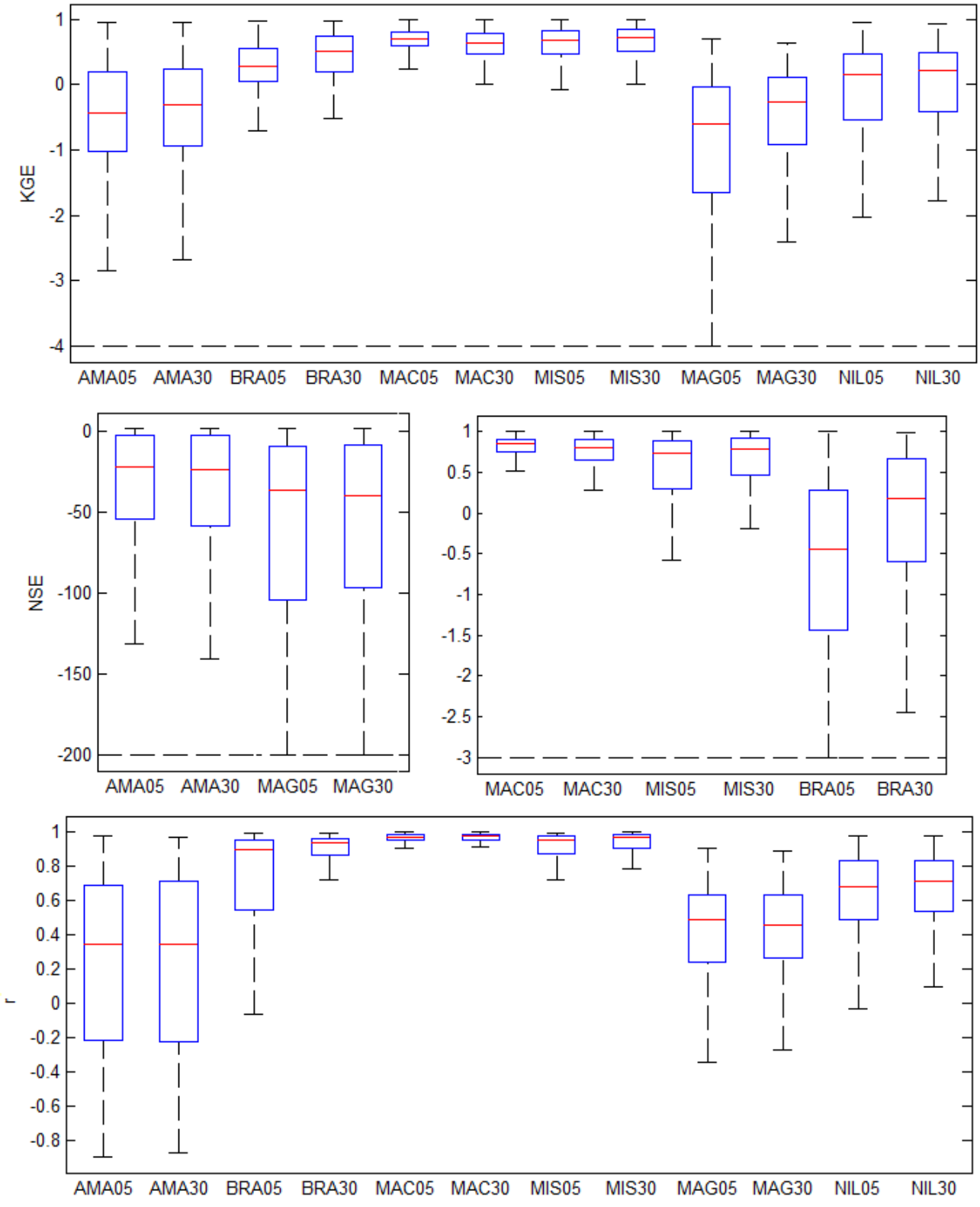


Figure A6: KGE, NSE and correlation score boxplots for all studied basins on both resolutions compared to the Fluxnet ET dataset. Note that the NSE boxplots have different y-axis. Outliers are not shown in the boxplots.

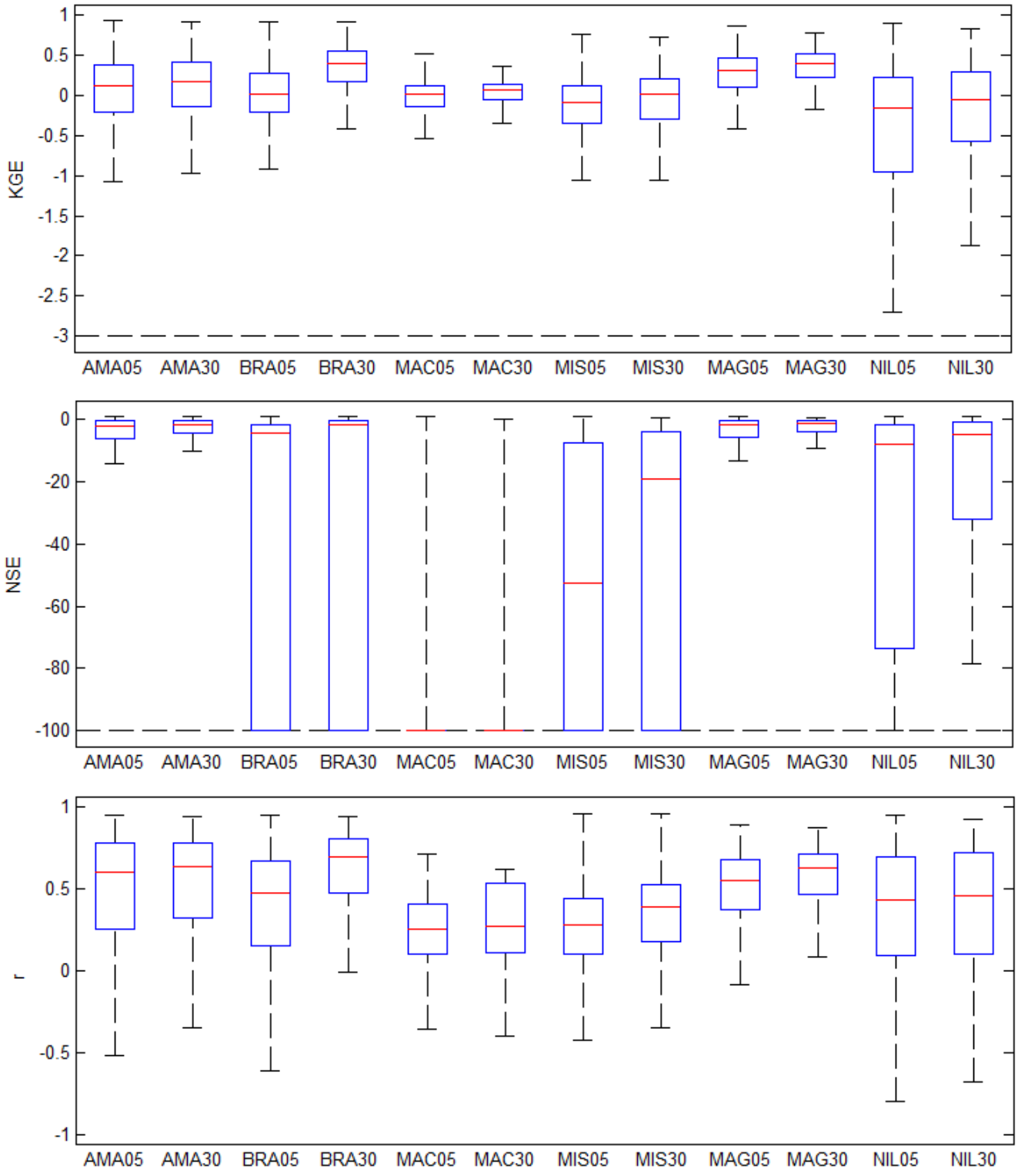


Figure A7: KGE, NSE and correlation scores boxplots for all studied basins on both resolutions compared to the ESA SM dataset. Outliers are not shown in the boxplots. NSE scores for the Mackenzie basins could not be calculated.

Table A8: Kling-Gupta, Nash-Sutcliffe efficiency scores and the coefficient of determination of the discharge at an observation point in the simulated river basins for the two simulations of PCR-GLOBWB. PCR-30 corresponds to the PCR-GLOBWB simulation on 30 arcminute resolution and PCR-05 is the PCR-GLOBWB simulation on 5 arcminute resolution.

		Kling-Gupta efficiency		Nash-Sutcliffe efficiency		r^2	
Basin *(location)*	Model (mode)	PCR-30	PCR-05	PCR-30	PCR-05	PCR-30	PCR-05
Amazon *(Jatuares)*		0.38	0.45	-1.80	-1.26	0.38	0.48
Brahmaputra *(Bahadurabad)*		0.43	0.56	0.55	0.62	0.83	0.73
Mackenzie *(Arctic Red River)*		0.35	0.43	0.46	0.52	0.67	0.82
Magdalena *(Calamar)*		0.72	0.59	0.41	0.00	0.70	0.81
Mississippi *(Vicksburg)*		0.35	0.46	-1.10	-0.35	0.47	0.43
Nile *(Dongola)*		-1.59	-1.65	-7.00	-7.12	0.26	0.61

Discussion: How sensitive is PCR-GLOBWB to changes in climate forcing, potential ET forcing and resolution?

The results have shown that the sensitivity of PCR-GLOBWB is not the same for each of these changes. Furthermore, there is a difference between the sensitivity on a local level and the accumulated difference in the form of the river discharge. On a local level the different precipitation input does not have a large influence, as the KGE, NS and correlation scores are similar for the CRU and GPCC simulations. A reason for this is the fact that the changes in precipitation are very small on a local level. The precipitation is increased on average 8.7% per cell (Müller Schmied et al., 2014), this results in relatively small increases in the total amount of water available and therefore the local effects remain small. However, the discharge is the accumulation of all the local water balances and therefore also influenced by the climate input. Differences between the discharge results are clearly visible: the average discharge of the GPCC simulations is higher for all the studied rivers. Only the Mississippi and the Nile have lower accuracy scores when the discharge is simulated with the GPCC precipitation in PCR-GLOBWB. The other rivers score higher when the GPCC simulations are compared to the CRU simulations. It is also worth noting that small differences in precipitation lead to larger discharge changes. Some basins are more sensitive to these changes than others, as the discharge changes are not the same for each basin, but all differences are above 1000 m^3/s.

PCR-GLOBWB is more sensitive to changes in ET forcing than to changes in the climate forcing. Changing the reference potential ET from the Hamon equation to the Penman-Monteith equation greatly increases the accuracy for the Amazon region. This effect is especially true for the ET itself as the accuracy scores show great improvement for the central basin, a region where the Hamon simulation scored lower. Not only the ET predictions, but also the soil moisture accuracy is improved. The increase is not as dramatic as the increase for ET, but the soil moisture scores are higher when the ET is forced by the Penman-Monteith equation. The discharge simulation for the Amazon shows a similar picture. It is more comparable to the observed discharge than any of the other PCR-GLOBWB simulations and the scores are more comparable to VIC-EB, which scores the highest for this basin.

The increase of the accuracy scores in the Penman-Monteith simulation indicate that the Hamon equation has trouble calculating an accurate PET, at least for the Amazon. This could be expected based on the equation. The Hamon equation is relatively simple and requires less information than the Penman-Monteith equation. It can be expected that, when more knowledge about an area is available and used to predict the PET, the quality of the PET prediction is better than that of an equation which uses less information. The results show that with a more accurate PET, PCR-GLOBWB also performs better in other areas of the waterbalance (the soil moisture and discharge). Further research could look into the effect that the Penman-Monteith has on other parts of the world.

The resolution results show that the accuracy of PCR-GLOBWB is different when the resolution is changed to 5 arcminutes. However, almost all accuracy scores are lower in the 5 resolution simulation. The actual reason for this is unknown and outside the scope of this thesis, but could be related to the comparison data. The discharge results show a different picture. Despite having worse local scores for the ET and soil moisture, the discharge has a higher accuracy in the 5 arcminute simulations, except for the Nile and the Magdalena.

A drawback of the resolution comparison is that the original resolution of the comparison data is not the same as the resolution of PCR-GLOBWB (5 arcminutes, 1/12th degree). The ESA soil moisture dataset and the ASMR-E SWE dataset have an original resolution of 15 arcminutes (1/4th degree) and the Fluxnet database of 30 arcminutes (1/2nd degree). This means that in order to compare both these datasets with the PCR-GLOBWB output they have to be upscaled to 5 arcminute resolution. The opposite is also true. Both the ESA and the ASMR-E data have to be downscaled for the comparison on 30 arcminute resolution. Moreover, some of the input data is not used on its standard resolution as well. An example is the climate forcing. This input has a resolution of 30 arcminutes and is upscaled during the simulation in PCR-GLOBWB itself. The result is that a discrepancy exists between the spatial patterns of precipitation and temperature in the model and the real world. This discrepancy also exists in the 30 arcminute simulations, but is less pronounced due to the size of each cell.

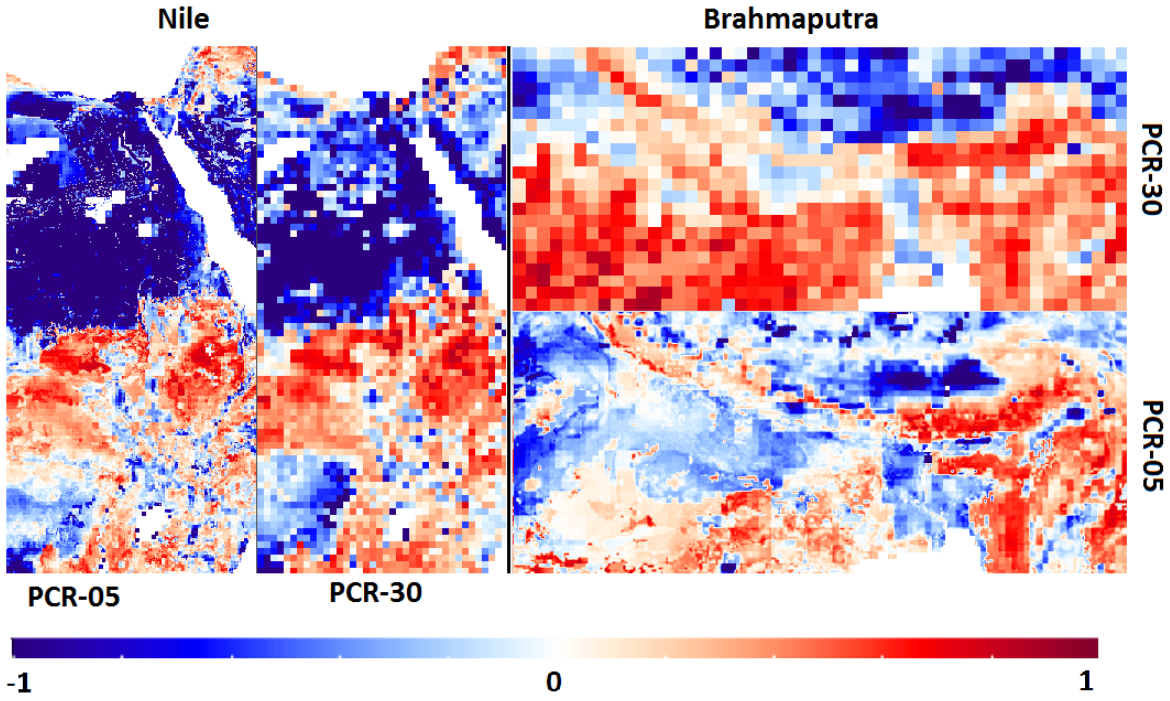


Figure A8: KGE scores for the Nile and Brahmaputra basin for the two different resolution versions of PCR-GLOBWB

The results would be similar if this was the only difference between the two versions of PCR-GLOBWB. This means that all results would look like the KGE scores for the soil moisture of the Southern Nile basin in Figure A8. Here both models produce somewhat similar results, with the same areas of high and low scores combined with a chessboard pattern on a more local level. This pattern indicates that due to the upscaling of the data, some cells correspond better to the data and some worse. The average of these cells is comparable to the cells in the PCR-30 version. However, this pattern is not visible for all basins. The boxplots in Figure A6 and Figure A7 already show that some basins perform worse in the 5 arcminute resolution and this can also be seen in the spatial distribution of KGE scores of Figure A8. The western parts of the Brahmaputra basin and the northern area of the Nile show a clear decrease of the KGE scores in the 5 arcminute simulation. The scaling issue is not the sole cause for this, as the area of the decrease is too large. This means that there are other issues in the model structure of PCR-GLOBWB that cause this decrease and reduce the accuracy of PCR-GLOBWB on a higher resolution.

B. Appendix 2: Supplementary figures

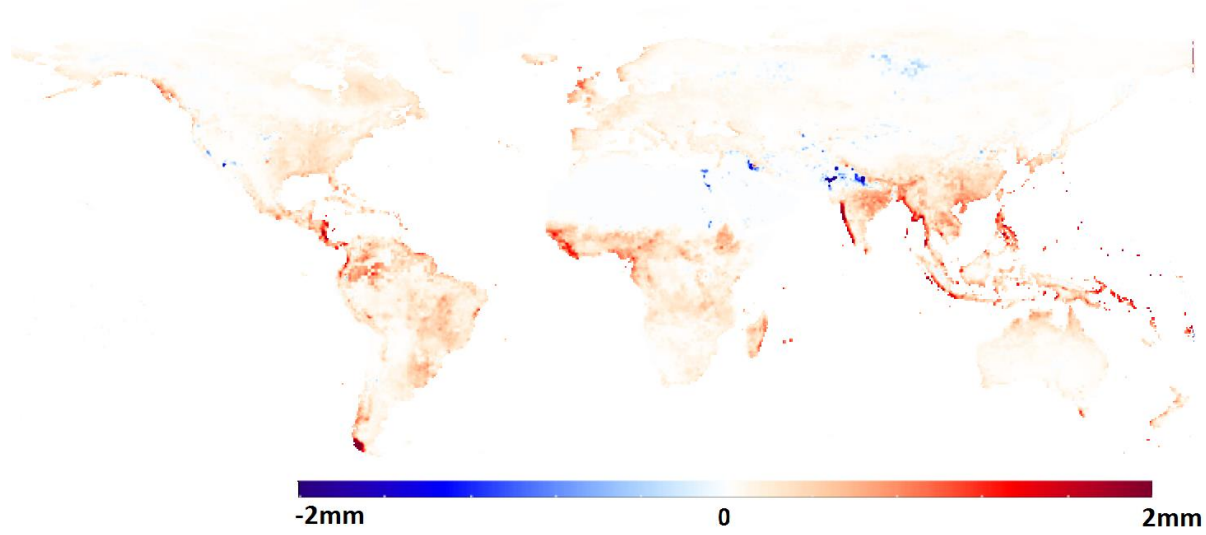


Figure B1: Difference of the average daily precipitation minus evapotranspiration rates as calculated by VIC-EB and VIC-WB (in mm/day) and averaged for the period 1979-2010. Positive values mean that the average daily precipitation minus evapotranspiration is larger in VIC-EB than in VIC-WB and vice versa. Note that this does not show accuracy, only the difference between the amount of water available after evapotranspiration for both cells.

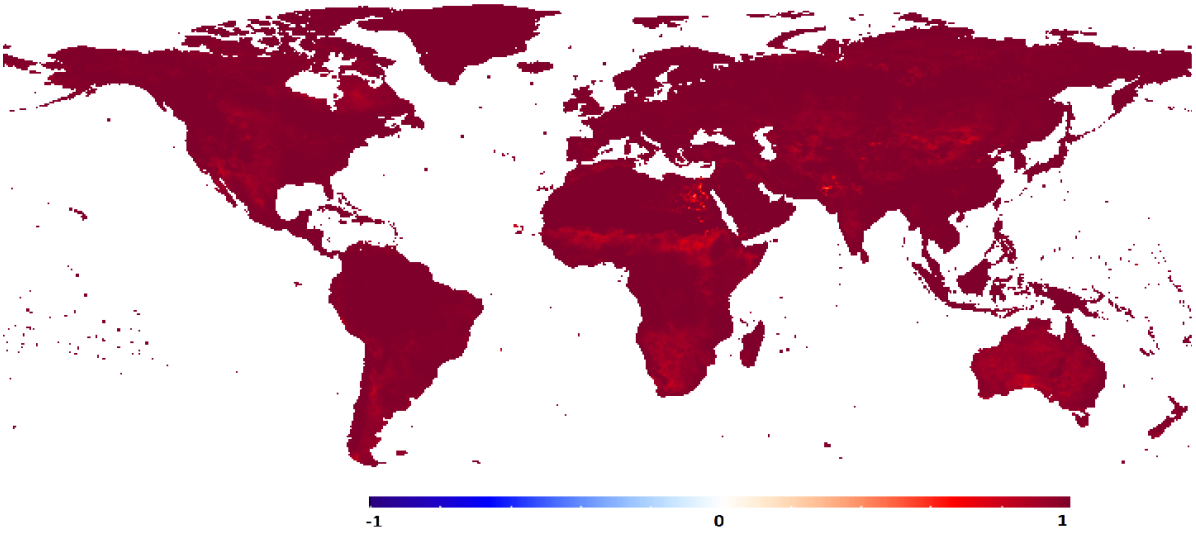


Figure B2: Correlation between the monthly averaged precipitation minus evaporation rates of VIC-EB and VIC-WB for the period 1979-2010.

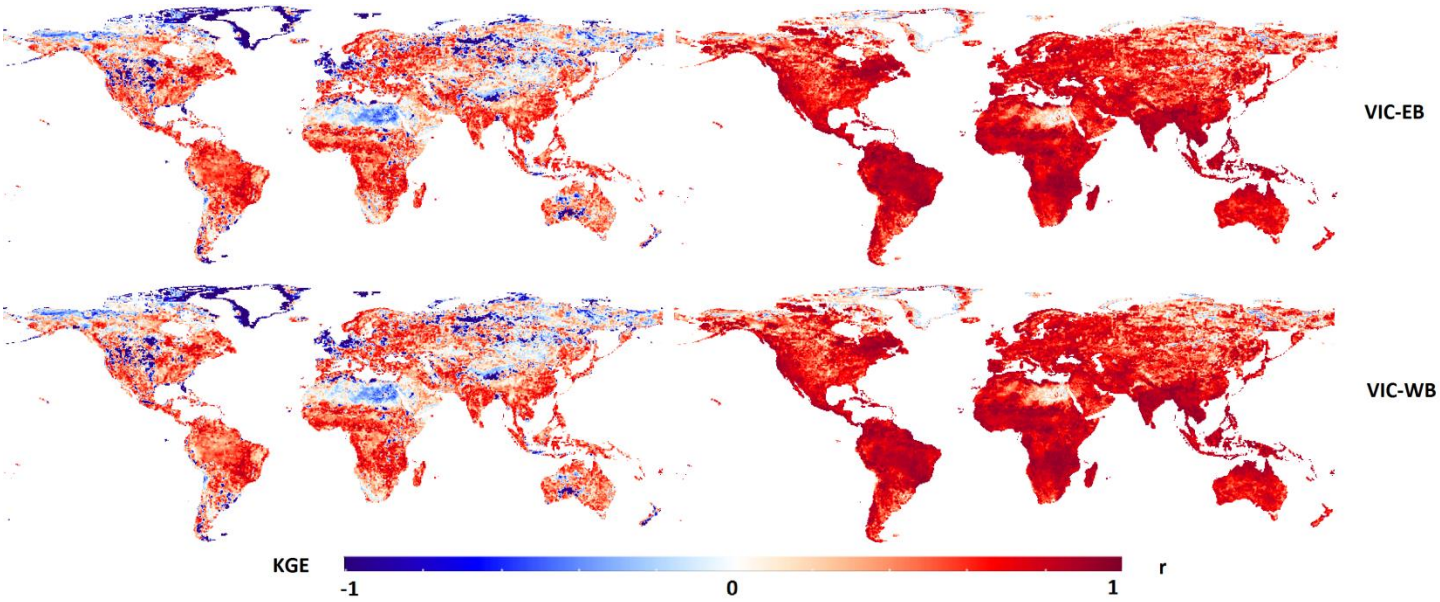


Figure B3: KGE and correlation scores of soil moisture output of VIC-EB and VIC-WB compared to the output of PCR-GLOBWB.