Utrecht Univsersity
Department of Physical Geography

Australian National University
Fenner school of Environment and Society

37.5 ECTS

Performance evaluation of large-scale hydrological models for different hydro-climatic zones in Australia

Author:
Nils Wagenaar

Student nr.
5574137

Supervisors:
dr. Rens van Beek
dr. Edwin Sutanudjaja
dr. ir. Geert Sterk
prof. dr. Albert van Dijk

January 17, 2018

Universiteit Utrecht

This master thesis is dedicated to my mother and brother who inspired and supported me during this period. I would also like to mention my dad, who is no longer of this world, for his inspiration and thorough support.

Abstract

Climate change will affect volumes and timings of river discharges and will therefore determine the distribution and amount of people experiencing floods and droughts in the future. Therefore, the need for reliable hydrological modeling increases in order to enhance water management and sustainable water use in the future. Outputs of these large-scale hydrological models have been compared with observations and it is seen that modeled output is not limited to input forcing only, which leads to uncertainty of modeled output. This research aims to fill the gap in existing literature by evaluating the performance of two global land surface models (LSM) and two global hydrological models (GHM) for actual evapotranspiration, soil moisture and runoff across different climate zones of Australia. This contribution offers an insight into the spatial performance distributions for different large-scale hydrological models for those hydrological variables. Also, this research evaluates the multi-model ensemble median for performance. The performances for the large-scale hydrological models are compared with performances from the calibrated national model, AWRA-L, which serves as benchmark in this research. For the evaluation of the large-scale hydrological models, the Tier-1 of the EU-funded EartH2Observe datasets have been used in this research. Furthermore, the Kling-Gupta Efficiency index is used as statistical metric for model performance.

For monthly runoff fields, W3RA performed best in the tropical climate zones ($KGE > 0.2 = 70\%, KGE > 0.7 = 25\%$ of all simulations). ORCHIDEE obtained the worst scores for this climate zone ($KGE > 0.2 = 0\%$). Both W3RA and and HTESSEL simulations resulted in the best performances in the arid climate zones ($KGE > 0.2 = 50\%$), while ORCHIDEE performed worst in this climate zone ($KGE > 0.2 = 0\%$). For the temperate climate zone, W3RA performed best ($KGE > 0.2 = 40\%$, $KGE > 0.7 = 5\%$) and AWRA-L worst ($KGE > 0.2 = 10\%$, $KGE > 0.7 = 3\%$). In addition, for runoff evaluation, there is no clear differentiation in performances for particular climate zones found between GHMs and LSMs. There were differences in model performance between the large-scale hydrological models, but this was not related to whether the model was a GHM or a LSM.

For monthly actual evapotranspiration simulations, the best performances for the tropical climate zones are obtained by HTESSEL and ORCHIDEE

($\overline{KGE}^{Tropical}_{HTESSEL} = 0.76$; $\overline{KGE}^{Tropical}_{ORCHIDEE} = 0.71$). The ensemble median obtained the lowest KGE values for this climate zone ($\overline{KGE}^{Tropical}_{Ensemble-median} = -0.1$). For the arid climate zones, PCR-GLOBWB performed best ($\overline{KGE}^{Arid}_{PCR-GLOBWB} = 0.51$), whereas ORCHIDEE simulations resulted in the lowest scores for this climate zone ($\overline{KGE}^{Arid}_{ORCHIDEE} = -0.28$). HTESSEL performed best for the temperate climate zones ($\overline{KGE}^{Temperate}_{HTESSEL} = 0.47$), whereas the lowest scores for the temperate climate zones are obtained by AWRA-L ($\overline{KGE}_{GHMs} = 0.21$). For actual evapotranspiration, GHMs perform on average better in arid climate zones ($\overline{KGE}_{GHMs} = 0.34$; $\overline{KGE}_{LSMs} = -0.412$), whereas land surface models obtain on average higher scores in the tropical climate zones ($\overline{KGE}_{GHMs} = 0.47$; $\overline{KGE}_{LSMs} = 0.74$).

After evaluating soil moisture fields, best performances are obtained by W3RA for the tropical climate zones ($\overline{KGE}^{Tropical}_{W3RA} = 0.42$). HTESSEL performed worst for this climate zone ($\overline{KGE}^{Tropical}_{HTESSEL} = -0.95$). The ensemble median performed best for both the arid and the temperate climate zones ($\overline{KGE}^{Arid}_{Ensemble-median} = 0.67$; $\overline{KGE}^{Temperate}_{Ensemble-median} = 0.48$), whereas AWRA-L performed worst in the arid climate zone ($\overline{KGE}^{Arid}_{AWRA-L} = 0.002$) and W3RA in the temperate climate zones ($\overline{KGE}^{Temperate}_{W3RA} = 0.04$). After evaluating soil moisture fields by large-scale hydrological models for performance, this research found no spatial differentiation in model performance between GHMs and LSMs. This research demonstrated that that the multi-model ensemble median leads to satisfying results for the three evaluated hydrological variables. However, it is not necessarily better than each individual model for all climate zones.

In order to decrease parameter uncertainty and increase reliability, PCR-GLOBWB has been calibrated using streamflow observations from four catchments in Australia. The effective parameters are K_{sat}, J and $StorCap$ and the calibrated parameter set is $f_j = 1.0, f_s = 1.0$ and $f_k = 0.25$. The calibrated PCR-GLOBWB model has been validated against streamflow records from all other catchments in Australia. This research found major improvements after validating the calibrated PCR-GLOBWB model with the reference scenario for the all climate zones. Taking all climate zones into account, the PCR-GLOBWB run with the default parameter setting obtains for 20% of its simulations a KGE > 0.2. Calibration increased

this percentage with 15% to 35%. Also, calibration of PCR-GLOBWB leads to an increase of 10% for simulations with KGE > 0.5. In addition, this research demonstrated that performance improvements differ in magnitude between climate zones ($Calibrated\ PCR - GLOBWB : KGE_{tropics} > 0.2 = 45\%$, $Reference\ PCR - GLOBWB : KGE_{tropics} > 0.2 = 10\%$; $Calibrated\ PCR{-}GLOBWB : KGE_{arid} > 0.2 = 15{-}20\%$, $Reference\ PCR{-}GLOBWB : KGE_{arid} > 0.2 = 10\%$; $Calibrated\ PCR - GLOBWB : KGE_{temperate} > 0.2 = 35\%$, $Reference\ PCR{-}GLOBWB : KGE_{temperate} > 0.2 = 25\%$).

Still, after calibration, performances changed from really bad KGE to bad KGE for some climate zones. This is mainly attributable either to bad model structure, poor forcing dataset or the wrong selection of effective parameters for this climate zone rather than the use of a sub-optimal combination of parameters in the PCR-GLOBWB model. Also, this research demonstrated that the calibration of PCR-GLOBWB using streamflow data negatively influences the performance for actual evapotranspiration fields by the calibrated PCR-GLOBWB model. Therefore, before calibrating a hydrological model and applying it to a certain region, the purpose of the modeling needs to be fully known. This research proved that global scale hydrological modeling could be a valuable source of knowledge for developing countries without a fine resolution hydrological model. However, further research needs to be carried out for both step-wise calibration to enhance applicability and for the improvement of forcing data at 0.5° or at smaller spatial resolution in order to enhance reliability of large-scale hydrological models.

Keywords: Physical Geography; Hydrology; Earth2Observe; Water Resources Reanalysis; Global Hydrological Models and Land Surface Models; Parameterization

Acknowledgements

First of all, I would like to express gratitude to my supervisors Rens van Beek, Edwin Sutanudjaja and Geert Sterk for their continuous support, enthusiasm, knowledge, patience and contribution to this thesis. Their guidance have assisted me in both the research process and writing the thesis itself. I would also like to mention Albert van Dijk, who assisted and supported me during my exchange at Australias National University (ANU). Lastly, I thank my family: My mother Brenda Renssen, my brother Joris Wagenaar and my girlfriend Roos Kallasingh, who supported me continuously throughout the thesis process.

Contents

List of Figures

List of Tables

1 Introduction

According to the UN comprehensive assessment of the Freshwater Resources of the World (WMO, 1997), more than two-third of the global population will experience water stress by the year of 2025. The increase of greenhouse gas concentrations in the atmosphere will result in climate change and this will affect volumes and timings of rivers discharges and groundwater recharges (Arnell, 2004). In fact, the growing knowledge about the interdependence of various earth systems has led to the need for integration of those systems in global simulation models (Wilby & Dessai, 2010). As temperatures will increase in the next decades, the atmospheres water holding capacity will increase and therefore more extreme hydrological events are likely to occur (Trenberth, 1999). Consequently, climate change will determine the amount and distribution of people experiencing droughts and floods in the future. For this reason, the need for reliable large-scale hydrological modeling is increasing in order to respond to these expected changes in the global hydrological cycle and global water resources. In fact, there is a need for water assessments on a regional/global scale to enhance water management and sustainable water use in the future. This growing demand for large-scale hydrological modeling has led to the EU-funded EartH2Observe project (EartH2Observe, 2015). This project aimed to construct a consistent 30-years water resources re-analysis dataset, which allows for enhanced insights on both the existing pressures on water resources and on the full extent of water availability globally.

There are two groups of hydrological models available: water balance models and land surface models. The major difference between these two model groups is their approach for evapotranspiration calculation. Evapotranspiration is the link between the energy balance and the water balance. Water balance model operating at the contintental/global scale are often referred as global hydrological models (GHMs).

On the one hand, water balance models are mostly conceptually-based distributed models and these models solely solve the terrestrial water balance within a catchment. These models are forced with prescribed meteorological conditions like temperature, precipitation and either net radiation or reference potential evapotranspiration as input (e.g. SWBM, PCR-GLOBWB) (Orth & Seneviratne, 2015 ; van Beek & Bierkens, 2009). Reference potential evapotranspiration can either be prescribed or calculated by the use of temperature and day length (Hamon, 1961). Thereafter, crop factors convert reference potential evapotranspiration into potential evapotranspiration.

On the other hand, Land surface models solve both the energy balance and the terrestrial water balance. The land surface energy fluxes are calculated by the use of temperature and vegetation cover from remote sensing (Wang et al., 2006).

Different water balance models and large-scale land surface models have already been developed. For instance, PCR-GLOBWB (Sutanudjaja et al., 2016 ; van Beek et al., 2011; Wada et al., 2014), LISFLOOD (Van Der Knijff et al., 2010), SWBM (Orth and Seneviratne, 2013) and WaterGAP3 (Flörke et al., 2013; Dll et al., 2009) are the main global hydrological models (GHMs) and ORCHIDEE (Krinner et al., 2005 ; Ngo-Duc et al., 2007 ; dOrgeval et al., 2008), JULES (Best et al., 2011; Clark et al., 2011), HTESSEL-CaMa (Balsamo et al., 2009), SURFEX-TRIP (Decharme et al., 2010, 2013) are the main global land surface models (LSMs). Next to that, the model W3RA is a hybrid model between a water balance and a land surface model.

These large-scale hydrological models have been run extensively throughout the last decades. Modeled output has been compared with available data and it is proven that modeled output is not only limited to input forcing (Sood & Smakhtin, 2015). So, for equal catchments, different models may result in different outcomes, which leads to uncertainty in model prediction. In fact, uncertainty arises from several factors including: input data, parameters, model structure and observational errors (Kauffeldt et al., 2016). Earlier studies have shown that uncertainty from model parameters and model structure can be substantial (Haddeland et al., 2011 ; Walker et al., 2003). Haddeland et al. (2011) found major differences in global/regional water fluxes and storage terms for an ensemble of 11 large-scale hydrological models, which used the same forcing dataset. Substantial differences were especially found in the partitioning between evapotranspiration and runoff, which caused major differences in runoff estimates by the large-scale hydrological models. In addition, a model inter-comparison project (MIP) between multiple land surface schemes has been carried out and demonstrated that land surface model simulations of streamflow and land surface - atmosphere fluxes hold large variability (Wood et al., 1998 ; Lohmann et al., 2004). Beck et al. (2016) evaluated streamflow simulations by 10 state-of-the-art large-scale hydrological models and found substantial differences in model performances for different hydro-climatic zones across the globe. Gudmundsson (2012) found large differences in the performances of large-scale hydrological models for different hydro-climatic zones, but demonstrated that, on average, the ensemble mean of model simulations outperformed most individual models. Moreover, the research showed that the ensemble mean performed consistently good for all hydro-climatic zones (Gudmundsson, 2012). In fact, multiple studies and model intercomparison projects (MIP) have found that the multi-model ensemble mean is generally superior to the results of any individual model and as good or even better than the best model at each point and time (Dirmeyer et al., 2006 ; Murphy et al., 2004). Still, little research has been conducted in this area.

This research aims to fill this gap in the existing literature by evaluating simulations of large-scale hydrological models over different climate zones. As such, this research is dedicated to obtain insights in how the spatial distributions of performances among different hydro-climatic zones are for different large-scale hydrological models. In addition, this research aims to investigate if the performances of the evaluated models could be related to whether these models are land surface models or water balance models. Australia has been selected for the evaluation of the large-scale hydrological models as it holds many different hydro-climatic zones (Table 1, Fig. 1). In this research two GHMs and two LSMs from the Tier-1 dataset of the EartH2Observe project (2015) will be evaluated for performance. In addition, the performance of the multi-model ensemble median will be analyzed as well.

Tab. 1: Climate codes descriptions for all climate zones in Australia based on Köppens climate classification. The first letter of the climate codes indicate the main climate group, which er A (tropical), B (arid) and C (temperate). Also, the green, blue and red colors correspond to tropical, arid and temperate climates respectively.

Climate codes	Climate descriptions
Am	Tropical monsoon climate
Aw	Tropical wet-dry climate
Af	Wet equatorial climate
BWh	Tropical and subtropical desert climate
BSh	Mid latitude steppe and desert climate
BSk	Tropical and subtropical steppe climate
Cfa, Cwa	Humid subtropical climate
Cfb	Marine west coast climate
Csa, Csb	Mediterranean climate

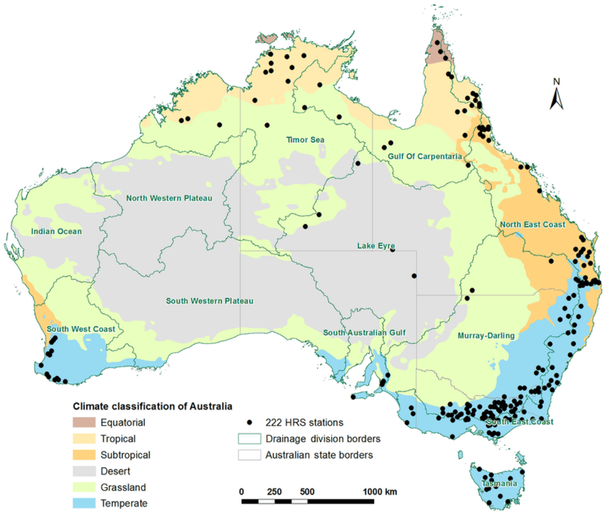


Fig. 1: Spatial distribution of hydro-climatic zones in Australia, major drainage divisions and hydrologic reference stations for streamflow (BoM, 2015)

The selected Tier-1 EartH2Observe large-scale hydrological models for evaluation are: PCR-GLOBWB (GHM), W3RA (GHM), HTESSEL-CaMa (LSM) and ORCHIDEE (LSM). For the Tier-1 water resources re-analysis, these models were all uncalibrated and at equal spatial resolution (0.5°). Thus, these large-scale hydrological models are suitable for direct comparison. In order to put the results into perspective, a comparison between the performances of the large-scale hydrological models and the national AWRA-L model will be performed as well. So, the AWRA-L model serves as benchmark for the evaluation of the large-scale hydrological models and is at 0.05° spatial resolution. Also, AWRA-L is calibrated with streamflow observations from Australia. The performance evaluation will be done for runoff, upper-layer soil moisture and actual evapotranpiration.

Due to the complexity of physically-based hydrological models, parameters are often highly uncertain and tend to loose their physical meaning (Beven, 1993). Calibration, often referred as parameter optimization (Simunek et al. 2012), is the process of optimizing unknown parameter values in order to decrease models predictive uncertainty and to improve model accuracy. In fact, the calibration of a hydrological model, where the parameters of a generalized hydrological model is adjusted, leads to a better representation of hydrological processes. In fact it reduces the models parameter uncertainty and increases the models reliability. Various studies have been carried out to demonstrate the importance of calibration on the performance of hydrological models (Nijssen et al. 2003; Duan et al. 2006). After, the "calibrated" model needs to be validated in order to demonstrate that the model is able to generate accurate simulations in different modeling circumstances. For this reason, the effect of calibrating the PCR-GLOBWB model on runoff estimates across various hydro-climatic zones of Australia will be examined as well. The above stated problem definition leads to the following research question, which is twofold:

- What is the performance of the four selected uncalibrated large-scale hydrological models for different hydro-climatic zones in Australia after evaluating runoff, soil moisture and actual evapotranspirtion simulations and how do these models perform compared to both the upscaled national hydrological model (AWRA-L) and the ensemble median?

- What is the performance of runoff estimates generated by the PCR-GLOBWB model after calibrating and validating the model using Australian observational streamflow data?

2 Methodology

2.1 Model selection

As mentioned earlier, 2 land surface models and 2 global hydrological models will be evaluated. For this study, the uncalibrated large-scale model simulations from tier-1 Earth2Observe dataset has been selected in order to compare the performances of these large-scale models fairly. Furthermore, the national AWRA-L model will be evaluated and serves as benchmark. One should keep in mind that AWRA-L is a calibrated model and is forced with a different dataset than the large-scale hydrological models (BAWAP vs. WFDEI). For this thesis, the models W3RA and PCR-GLOBWB are used as global hydrological models and the models HTESSEL-CaMa and ORCHIDEE as land surface models. The GHMs are originally developed in order to simulate (sub)-surface water storages and fluxes, while LSMs focus more on the interactions between soil, water and atmosphere in climate models (Bierkens, 2015). Both GHMs and LSMs solve the water balance. However, LSMs solve both the water and the energy balance. This leads to the potential to estimate hydrological partitioning more accurately. LSMs generally have a more complex model structure and contain more parameters than water balance models. As a result, due to their complexity and large number of parameters, LSMs are often not calibrated. GHMs have daily temporal resolution, whereas LSMs generally have sub-daily temporal resolution. The main reason for this is that LSMs aims to capture the diurnal cycle of evapotranspiration processes. Spatial resolution is 0.5° for all selected large-scale hydrological models. In general, LSMs include more soil and snow layers than GHMs as these models are more complex (Beck et al., 2016). The use of more soil layers in LSMs is due to parameterization of soil processes (e.g. heat fluxes in soil layers (Liang et al., 1999)).

In the next subsection (Sect. 2.2), the different models used for performance evaluation will be described. Sect. 2.3 explains the general model inter-comparison framework. Sect. 2.4 is dedicated to the description of the observational datasets, which will be used for the large-scale hydrological model evaluation.

2.2 Model description

Table 2 summarizes the major similarities and differences between model characteristics of the selected Tier-1 GHMs and LSMs. This Table serves as a reference for the next part of this thesis, where major features and general concepts of the selected large- scale hydrological models are described individually.

In the following subsections we will discuss the most important models in more detail. We start with the PCR-GLOBWB global hydrological model, followed by the W3RA model, whereafter the HTESSEL-CaMa model is discussed. Then, the ORCHIDEE model is described, and finally the AWRA-L model is explained.

2.2.1 PCR-GLOBWB

The first global hydrological model is the PCR-GLOBWB (PCRaster Global Water Balance) global hydrological model. This model is built for regional and global purposes and has been developed at the department of physical geography of Utrecht University. The PCR-GLOBWB model represent the terrestrial hydrology on a grid, which has currently a spatial resolution of a 0.5°. The temporal resolution is daily. In fact, on each grid cell, the model uses process-based equations in order to determine the soil moisture storage in 2 soil layers. At the same time, the exchange of water between land surface and atmosphere is calculated for each grid-cell. The exchange of water between land surface and atmosphere is driven by processes like for example precipitation, snow accumulation/melt and evapotranspiration. Processes responsible for water exchange between the vertical soil layers are percolation and capillary rise. PCR-GLOBWB calculates river discharge by means of accumulating runoff per grid cell, which is routed along the drainage network. The equations for routing are based on the kinematic wave approximation of the Saint-Venant equation with the momentum equation based on Mannings equations. However, routing of river flow can be also be calculated by the computationally efficient travel time approach (Deursen, 1995). PCR-GLOBWB is forced with temperature and precipitation from general circulation models. Potential evapotranspiration can either be forced or calculated by temperature and day length (Hamon, 1961). The variability

Tab. 2: Model characteristics for the selected tier-1 EartH2Observe models and the national AWRA-L model

Model name	PCR-GLOBWB	HTESSEL-CaMa	ORCHIDEE	W3RA	AWRA-L
Model type	GHM	LSM	LSM	Hybrid GHM/LSM	LSM
Evapotranspiration	Hamon (1961)	Penman-Monteith	Bulck ET$_p$, Barella -Ortiz et al. (2013)	Penman-Monteith	Penman-Monteith
Snow	Temperature based meltfactor,	1 layer,	1 moisture layer, 1-5 thermodynamic layers,	Mass balance	Not included
Soil	2 layers	4 layers	11 layers	1 layer	3 layers
Runoff	Saturation excess	Saturation excess	Green-Ampt infiltration, gravitational drainage	Saturation, infiltration excess	Saturation, infiltration excess
Reservoirs included	No	No	No	Yes	Yes
Lakes included	Yes (GLWD dataset)	No	No	Yes	Yes, BoM Water Storages
Water use included	Not in tier-1	No	Irrigation only	No	Yes
Routing	Kinematic wave with floodplains	Ca-Ma Flood River channel, floodplains, local inertial equations	Linear cascade of reservoirs at subgrid level	Kinematic wave	AWRA-R model
Timestep	1 day	1 hour	5min energy balance, 3hr routing	1 day	1 day
Resolution	0.5°	0.5°	0.5°	0.5°	0.05°
Forcing	WFDEI	WFDEI	WFDEI	WFDEI	BAWAP
References	Sutanudjaja et al. (2014), van Beek & Bierkens (2009)	Yamazaki et al. (2011)	d'Orgeval et al. (2008)	van Dijk et al. (2013)	Viney et al. (2015)

within a grid is processed by subdividing land cover in two classes: short and tall vegetation classes, which relies on the GLCC dataset (USGS EROS Data center 2002) Furthermore, PCR-GLOBWB developed a new version, which also includes a water demand module and which introduces reservoirs and irrigation areas (Sutanudjaja et al., 2014). However, this is not included in the tier-1 EartH2Observe simulations. Moreover, for tier-1 EartH2Observe, PCR-GLBOWB was not calibrated.

2.2.2 W3RA

The second global hydrological model is the W3RA model. In this model, the landscape component of the AWRA (Australian Water Resources Assessment) system (AWRA-L) has been used as basis for the development of the W3RA (World-Wide Water Resources Assessment) model (van Dijk, 2010a ; van Dijk, 2010b). The W3RA model is considered a hybrid model between a lumped catchment model and a simplified grid based land surface model. However, the model resembles more a global hydrological model rather than a land surface model. The AWRA-L model is not detailed in its description in order to make it applicable to regions where few ground observations are available, which is typical for Australia. Like PCR-GLOBWB, AWRA-L distinguishes two hydrological response units (HRU), which are deep-rooted tall forest and shallow rooted short vegetation. For each HRU, the vertical processes are described by: net radiation balance; partitioning of precipitation (van Dijk, 2001) and net precipitation (van Dijk, 2010c); water balance of three unsaturated soil layers; transpiration (Yebra et al., 2013); groundwater, surface and soil water evapotranspiration; vegetation canopy dynamics; groundwater dynamics (van Dijk, 2010c); surface water body dynamics (van Dijk, 2001). For the latest version of the model (v5.0), AWRA-L was calibrated against streamflow, actual evapotranspiration and soil moisture (Viney et al., 2015). The original AWRA-L v1.0 model was modified for the global application (W3RA), as the model lacked description of snow processes. For this purpose, the HBV96 snow model was implemented (Lindstrom et al., 1997). In this model, generated runoff propagates by means of kinematic wave approximation and it uses a global routing scheme with 0.5° flow direction grid (Oki et al., 2001). Moreover, W3RA streamflow estimates do not take anthropogenic influences into account (i.e. reservoirs, dams, abstraction). For the tier-1 EartH2Observe purpose, this model was not calibrated.

2.2.3 HTESSEL-CaMa

The third model is the HTESSEL-CaMa model. The first part consists of the HTESSEL (Hydrology in the Tiled ECMWF Scheme for Surface Exchanges over Land), which is a land surface model. HTESSEL calculates the response of the land surface due to atmospheric conditions. In fact, the surface water and energy fluxes are calculated as well as the temporal evolution of vegetation interception and snowpack conditions, soil moisture content and soil temperature. These grid cell calculations have been done independently, which means that there is no horizontal interaction between soil columns. For the computation of water and energy transfer, the soil has been subdivided into four layers. The Fourier law of diffusion is used for heat transfer in the soil, whereas the vertical movement of water in the unsaturated zone is captured by Richards equation combined with Darcys law. A variable infiltration rate, which incorporates sub-grid variability due to orographic differences, has been used for the computation of surface flow (Balsamo et al., 2009). Subsurface leaves the bottom soil layer as free drainage. The outcomes of the HTESSEL land surface model (surface and subsurface runoff) serves as input for the CaMa (Catchment-based Macro-scale) flood plain model. In this model, all river networks globally have been subdivided into hydrological units (discretized). This favors the efficient computation of flow at the global scale (Yamazaki et al., 2009). The simulations of HTESSEL-CaMa model were carried out at 0.5° by 0.5° spatial resolution. HTESSEL-CaMa used the default parameter setting for the tier-1 EartH2Observe project.

2.2.4 ORCHIDEE

The last model we are going to test is the ORCHIDEE. ORCHIDEE (ORganizing Carbon and Hydrology In Dynamic EcosystEms) is a land surface model, which is part of the IPSL (Institute Pierre Simon Laplace) earth system model. ORCHIDEE can run either coupled with the earth system model (IPSL-CM5) or stand-alone offline. For the purpose of EartH2Observe, ORCHIDEE works at three different scales: The model solves the energy balance at 0.5° spatial resolution, which is determined by the forcing data; The hydrological balance is solved separately at three different tiles on a grid box, which sizes depend on the vegetation distribution; The calculation of river flows through basins, which are defined at 0.5° by 0.5° resolution. A time-splitting procedure is used for the computation of partitioning between runoff and surface infiltrations, which allows a temporal resolution of <30min. The ORCHIDEE model distinguishes 13 different vegetation types, which are grouped in 3 ensembles (bare soil, grass/crops and trees). Interception loss and transpiration values are calculated for each individual vegetation type, whereas root uptake and through fall values are aggregated per vegetation group. This leads to three calculations for the hydrological balance at each tile on a grid box. The dataset provided by Reynolds et al. (2000) defined three different soil types, which are used in ORCHIDEE. The dominant soil type at a grid box is used for each tile. The river flow routing is described by Hagemann and Dumenil (1997) and Miller et al. (1994). Tier-1 EartH2Observe ORCHIDEE simulations were made by using the default parameter setting.

2.2.5 AWRA-L

The performances for the large-scale hydrological models will be compared with AWRA-L, which serves as benchmark in this research. The Bureau of Meteorology and the Commonwealth Scientific and Industrial Research Institute (CSIRO) have initiated the development of the Australian water resources assessment modelling system (AWRAMS). This modelling system consists of two parts: landscape modelling part (AWRA-L), which computes landscape water balance surface fluxes; river modelling part (AWRA-R), which aims to estimate river water balance fluxes. The simulations of these models are used as key information source for products as Water in Australia reports (WIA), climate briefings and national water accounts (NWA). The AWRA-L modelling part has a 0.05° spatial resolution and is a one-dimensional grid-based water balance model. The model represents groundwater, soil and surface water stores. The temporal resolution of the AWRA-L model is daily and the models outcomes are gridded estimates for evapotranspiration, runoff, deep drainage and soil moisture. Furthermore, the model operates at regional to global scale and has a temporal range of modeled data larger than 100 years (Hafeez et al., 2015). Vegetation is distinguished as deep and shallow rooted vegetation and within each grid-cell the water balance for both types of vegetation is calculated separately. ET$_{act}$, streamflow and catchment average soil moisture observational data have been used for the calibration procedure of the most recent version of AWRA-L (AWRA-L v5). Approximately 300 catchments were chosen for calibration and another 300 catchments were used for validation of the model. According to Frost et al. (2015), these catchment needed to satisfy the following conditions: >50 km for decent catchment representation on a grid-cell; unregulated streamflow; no irrigation and land use impacts; sufficient record length of observational data (>10yrs length). The calibration procedure used an objective function to optimize the statistical fit between modeled output and observations. Nash-Sutcliffe efficiency, correlation and monthly/daily bias were used as metrics for optimization (Frost et al., 2015).

2.3 General model intercomparison framework Earth2Observe

As mentioned before, two LSMs (ORCHIDEE and HTESSEL) and two GHMs (W3RA and PCR-GLOBWB) will be evaluated for performance. In this research, performance evaluation will be done for actual evapotranspiration, soil moisture and runoff (Table 3) between 01-01-2001 and 31-12-2011 at locations from available observational data (Sect. 2.4).

Tab. 3: Selected Earth2Observe output variables for performance evaluation

Long name	Units	Definition	Positive direction
Total evapotranspiration	$kg * m^{-2} * s^{-1}$	Sum of all evapotranspiration sources, averaged over the corresponding grid cell	Downwards
Total runoff	$kg * m^{-2} * s^{-1}$	Average of all liquid water draining from the land surface	into gridcell
Surface soil moisture	$kg * m^{-2}$	best of 5cm depth soil moisture or first layer	-

This research will evaluate both daily and monthly Tier-1 simulations from the selected large-scale hydrological models as monthly simulations generally tend to perform better than daily for large-scale hydrological models (Spruil et al., 2000). For runoff, this is mainly due to the inability of the large-scale hydrological models to capture extremes (e.g. peak flows) (Spruil et al., 2000 ; Mutenyo et al., 2013). Therefore, minimum observational record length is set to 2 years for the three hydrological variables in order to retrieve meaningful performance statistics from the model simulations. Once these model performances are collected, a spatial distribution of performances over the Australian continent will be made in order to investigate whether these performances are related to the combination of model structure and climatic conditions. As earlier studies demonstrated (Dirmeyer et al., 2006 ; Murphy et al., 2004), the multi-model ensemble mean is generally superior to each individual model and as good or even better than the best model at each point and time. This statement will be analyzed as well, but for the ensemble median instead of the ensemble mean as this excludes the severe effect of outliers on the statistic (Rodda & Little, 2015). Outcomes for performance evaluation of model simulations for actual evapotranspiration and soil moisture should be analyzed with care, as grid-based average values will be compared with point observations. Streamflow observations, in contrast, are the integrated hydrologic response of a catchment. For this reason, results obtained by evaluating streamflow simulations are more meaningful.

For AWRA-L, the evaluation period will be between 01-01-2005 and 31-12-2010 due to data availability and results obtained serves as benchmark for the performances of the large-scale hydrological models. As mentioned in section 2.2, the AWRA-L model is forced by BAWAP at 0.05° instead of WFDEI at 0.5°. Consequently, upscaling of the 0.05° resolution model outputs is needed for direct comparison with the large-scale hydrological models.

2.3.1 Upscaling AWRA-L model

There are various upscaling procedures suggested in literature (Qin et al., 2015), which are: Block kriging, Simple averaging, apparent thermal inertial (ATI) and hydrologic model based methods. This research uses the simple averaging method (Eq. 1).

$$\overline{B} = \frac{1}{n} * \sum_{i=1}^{n} a_i \tag{1}$$

Where a_i represent all fine resolution grid-cells inside one coarse resolution grid-cell. The major advantage of this re-sampling method is that this procedure is mass conservative. In contrast, the main interpolation methods are not mass conservative. This means that for upscaling by simple averaging, no mass in the system is lost. Upscaling by interpolation, in contrast, often leads to loss of mass in the system (Lagrava et al., 2012). As a result, mass leaving the fine grids does not equal the mass entering the coarse grid.

Therefore, the simple averaging method is preferred. Still, this method has its own disadvantage as aggregation by simple averaging often leads to errors and loss of information (Sablok & Aziz, 2008). Moreover, the nonlinear relationship the soil moisture state and associated physics as well as the heterogeneous soil moisture fields has led to increased interest in the problems associated with soil moisture aggregation (Crow & Wood, 2002). For this research, this should be kept in mind when comparing soil moisture fields with observations. Figure 2 illustrates the process of upscaling a fine resolution grid to a coarse resolution grid.

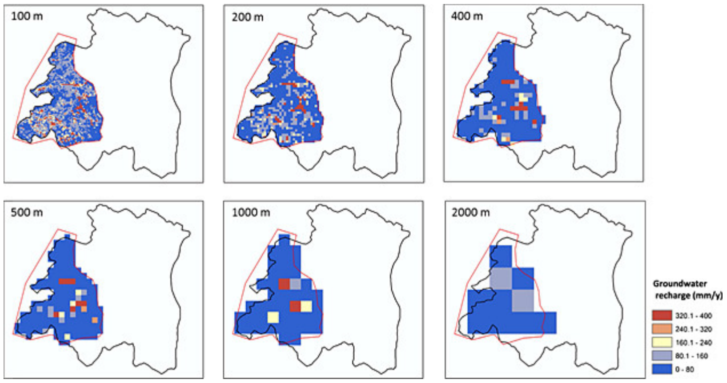


Fig. 2: Upscaling process from a fine resolution grid to a coarse resolution grid. source: He et al., 2015

As mentioned in the introduction (Sect. 1), model calibration/validation is an important step in order to adjust a generalized model to local/site specific processes and conditions in order to improve model accuracy and reduce model predictive uncertainty. In the next subsection, the calibration/validation procedure will be explained.

2.3.2 Calibration and validation procedure

Brute-force calibration Hydrological model output uncertainty is caused by input forcing, model parameters, model structure and observational errors. Calibration of a hydrological model is an essential part of modeling as it explores effects of different parameter combinations on model output. Furthermore, calibration searches for the optimal parameter combination, which reduces parameter uncertainty and increases reliability of the model. The evaluation part for runoff of the uncalibrated PCR-GLOBWB model is used as information source to tune parameters in the brute-force calibration procedure. The first step of this procedure is to identify parameters, which possibly improve the model performance for runoff. For this part, we will look at both the simulated and observed hydrographs. When these possible effective parameters are identified, a large number of runs will be simulated and evaluated using an objective function. For this calibration step, four suitable catchments will be selected and different streamflow runs for these catchments will be compared with observational data. Calibration will be based on monthly average streamflow simulations (m/d). The selected catchments need to satisfy the following criteria for inclusion in this calibration step:

- Catchment area should be sufficiently large compared to models grid-cell resolution (i.e. 0.5° resolution), to ensure catchment size is representative for 0.5 ° resolution (Beck et al., 2015); Shrestha et al., 2006)
- Inclusion of catchments located in different climate zones for the calibration step. This leads to better possibilities for regionalization of model parameters to other catchments in Australia.
- The simulations for potential catchments have to perform sufficiently (Tier-1 Earth2Observe performance, Sect. 3.1.1, based on a selected statistical performance metric, Sect. 2.3.3) to ensure the model structure is able to represent the hydrological processes for that catchment.
- Sufficient observational data record length, > 3 years, is required in order to obtain consistent and stable parameter values (Li et al., 2010).

Information regarding these criteria will be gathered from the evaluation of the tier-1 EartH2Observe PCR-GLOBWB model run. Calibration of the PCR-GLOBWB model will be done from 01-01-2001 till 31-01-2011.

Once these catchments are selected, multiple runs of the PCR-GLOBWB model will be executed with different parameter combinations. This research will perform a global calibration procedure, in which one parameter set is sought for all catchments (Gaborit et al., 2015). Based on the concept of equifinality (Beven & Binley, 1992), parameter combinations leading to satisfying results are called behavorial parameter sets. Lu et al. (2014) and Hamraz et al. (2015) defined behavorial parameter sets as the top 1% best performing parameter combinations based on a likelihood function. This study defines the behavorial range as the top 5% of best performing parameter combinations based on a predefined objective function due to the limited amount of runs in the brute-force calibration procedure.

Validation After we have chosen the ”global” optimal parameter set for the selected catchments, validation of the calibrated PCR-GLOBWB model will be done for all other catchments. For the validation of the calibrated PCR-GLOBWB model, we compare the performance of the calibrated PCR-GLOBWB model with the uncalibrated PCR-GLOBWB model, which was used for the EartH2Observe project. Also, performance of the calibrated PCR-GLOBWB model will be tested against the performances of the benchmark AWRA-L model, and the ensemble median of the large-scale hydrological models. Validation is essential as problems may emerge when testing the calibrated model at other catchments with different hydrological conditions. Often, performances obtained for these catchments are less satisfying compared to the performances obtained from catchments used for the calibration of the hydrological model. This problem is often referred to as a ”regionalization problem” as calibrated parameters for a certain hydrological situation may not apply for other regions. Especially in this research, where many catchments across different climate zones in Australia are included, unsatisfying results may arise. So, for this reason it is very important to validate the calibrated model.

2.3.3 Functions and definitions for model evaluation

Performances of the selected large-scale hydrological models (Tier-1 model output) for runoff and actual evapotranspiration will be quantified by the Kling-Gupta Efficiency index (KGE). Moreover, the calibration of PCR-GLOBWB will be performed using the KGE as objective function. Three statistical metrics are combined , resulting in the KGE index (Eq. 5). These metrics are bias (Eq. 2), Pearsons correlation coefficient (Eq. 3) and variablity ratio (Eq. 4).

$$Bias = \sum_{i=1}^{n} \frac{Simulated_i}{Observed_i} \tag{2}$$

$$Pearsons\,Correlation\,Coefficient = \frac{n\sum X_i Y_i - \sum X_i \sum Y_i}{\sqrt{n\sum X_i^2 - (\sum X_i)^2} - \sqrt{n\sum Y_i^2 - (\sum Y_i)^2}} \tag{3}$$

$$Variability\,ratio = \frac{\sigma_s}{\sigma_o} \tag{4}$$

$$Kling - Gupta\,Efficiency = 1 - \sqrt{(Correlation - 1)^2 + (Bias - 1)^2 + (Variability - 1)^2} \tag{5}$$

In fact, the KGE represents the Euclidian distance between the three statistical measures and an ideal point in a 3-dimensional space. For the KGE, the ideal point in the 3-dimensional space is (1, 1, 1). So, minimizing the the Euclidian distance leads to higher KGE scores. The value for KGE ranges from $-\infty$ to 1 where 1 means a perfect fit between model simulations and observations. This means, the higher the values for the three statistical measures, the smaller the Euclidian distance and the better the KGE value (Eq. 5). This study uses a modified KGE for soil moisture evaluation. Soil moisture is calculated over different depths between models and measured over different depths between different observational soil moisture networks. Also, the different networks have different sampling resolution and different uncertainties (Sect. 2.4.3). Therefore, the bias term (Eq. 2) in the KGE equation (Eq. 5) is omitted. Now, the KGE value corresponds to the Euclidian distance from an between the coordinates for correlation and variance ratio to an ideal point (2-dimensional). The value for the modified KGE still ranges between $-\infty$ and 1. For both the default KGE values as the modified KGE values, a classification for performance is made in Table 4.

Tab. 4: Classification of KGE values for model evaluation

Classification	KGE score
Very good	$KGE > 0.7$
Good	$0.5 < KGE < 0.7$
Satisfying	$0.2 < KGE < 0.5$
Unsatisfying	$KGE < 0.2$

During the calibration of the PCR-GLOBWB model, several PCR-GLOBWB runs will be evaluated for performance at the selected catchments (Sect.

2.3.2). A function (Eq. 6) is defined to exclude the unproportional effect of very bad performances for certain catchments:

$$\overline{KGE} = \frac{\sum_{i=1}^{n} max(0, KGE_i)}{n} \qquad (6)$$

Eq. 6 shows that for each calibration run, the average of the maximum value between zero and the KGE value for certain catchments will be evaluated. This is done as for some areas the model might perform very bad. However, the difference between bad and very bad is not related to a sub-optimal parameter set.

2.4 Observations

in this subsection, the observational datasets used for the evaluation of the Tier-1 Earth2Observe large-scale hydrological models will be described. Firstly, in section 2.4.1, we demonstrate which dataset is used for runoff comparison. Then, in section 2.4.2, a description of the observational actual evapotranspiration datatset is provided. Lastly, the soil moisture dataset will be described in section 2.4.3.

2.4.1 Runoff

A previous study (Zhang et al., 2013) has collected data for runoff throughout Australia. This datasets consists of runoff data from catchments and has been used for calibration and validation of the AWRA-L model. This dataset is made taking several criteria into account:

- Sufficient data record length (> 2 years)
- Covering recent times
- Good data quality, which means not too many observations far in the past (where streamflow records are potentially of lower quality (Zhang et al., 2013)).
- Catchments are larger than 50km^2 in size as the widely used input for

meteorological forcing is at a 50km^2 resolution.

- Catchments need to be unimpaired/unregulated catchments as most hydrological models simulate natural flow (Zhang et al., 2013). For the Tier-1 Earth2Observe, this arguments holds as well as the majority of the selected large-scale models for evaluation did not include anthropogenic influences (e.g. dams, reservoirs and lakes) in their model (Sect. 2.2).

Having a good quality dataset for runoff from unimpaired catchments is extremely important for benchmarking hydrological models. This is important as these hydrological models are used for water resources assessments on a large-scale in Australia. This dataset is obtained by state water agencies and quality checks were undertaken after. This has resulted in a dataset consisting of 780 unregulated catchments across different climate zones of Australia. For this research, no further selection for catchment suitability has been undertaken for the following reasons:

- To cover substantial hydro-climatic zones with a substantial sample size
- To avoid the steering of research in a certain direction for satisfying results
- To investigate whether global modelling is applicable to all catchment scales

The observational dataset consists of daily streamflow values at the catchment outlet (mm/d). This value for runoff is the integrated response of hydrological processes within the catchment. The Tier-1 EartH2Observe datasets from the selected large-scale hydrological models hold runoff values for each 0.5° grid ($kg * m^{-2} * s^{-1}$). The weighted average spatial aggregation is used in order to come up with runoff estimates for large-scale hydrological models at the outlet. For this approach, runoff estimates for grid-cells corresponding to a certain catchment will be summed. After, this value is divided by the number of corresponding grid-cells covering the catchment to come up with a average runoff estimate. This process will be repeated for each time-step. In literature, it is a general opinion that in most cases this method provides closest estimates for runoff (Sauquet et al., 2000). Fig.

3 shows the distribution of unimpaired catchments throughout Australia. These catchments will be evaluated in this study.

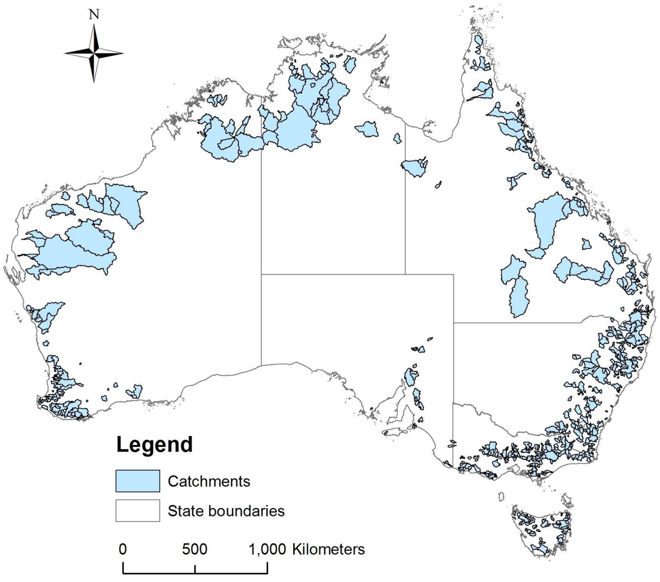


Fig. 3: Spatial coverage of unimpaired catchment across Australia. Source: Zhang et al. 2013

2.4.2 Actual Evapotranspiration

Data for evapotranspiration is obtained by the Ozflux national ecosystem research network. Ozflux uses micro-meteorological flux stations, which uses the eddy-covariance statistical method to estimate the exchange of water vapor, heat, carbon dioxide and methane between the land surface and atmosphere. The data contains daily values for evapotranspiration in $MJ * m^{-2} * d^{-1}$ for 18 fluxtowers. These values will be divided by the latent heat of vaporation $(2.45MJ*kg^{-1})$, because modeled data is in $kg*m^{-2}*s^{-1}$. Ozflux evapotranspiration observations are available between 01-01-2001 and 31-12-2011. As mentioned in section 2.3), a sufficient ($>$ 2 years) dataset record length is required for meaningful statistical performance analysis.

2.4.3 Soil Moisture

This study uses soil moisture data equal to the dataset used in Holgate et al. (2016). This dataset consists of soil moisture measurements starting from 01-01-2001 till 31-12-2014. Again, sufficient data record length was required for statistical robustness. Soil moisture is measured by three networks, which in turn use different measurement techniques. All three products delivered their data in volumetric soil moisture fraction (m^3/m^3).

Ozflux The first network is Ozflux. As already mentioned, Ozflux is part of a global network with >500 micro-meteorological stations where exchanges of water vapor, energy and carbon are continuously measured. There are 37 stations in Australia with 30 currently active stations. Soil moisture profiles are measured by frequency domain reflectometers for every 30 minutes. Provided soil moisture data are in volumetric units or fraction and the measurements were performed over the upper 10cm topsoil. Soil moisture data from these stations were made available for 22 of the 37 stations. Data from these stations are used in this study

Oznet Oznet contains a network of measurement sites comprising an area of 82000km^2 in the Murrumbidgee catchment in southeastern Australia. The sites collect data for rainfall, soil moisture and soil temperature at 20 to 30 min time resolution. These measurement sites are operationally since 2001 and collected data primarily for the root-zone (up to 90 cm). After, topsoil measurements were carried out as well (0-5cm). The initial stations carried out their measurements using water content reflectometers. Temperature and soil type information were used to convert these measurements to volumetric water content. Later, other measurement techniques were used as well in which volumetric moisture content is inferred from measured conductivity and the dielectric constant. Oznet data from 48 stations were made available. These are included in this research

CosmOz The third soil moisture measurement network operational in Australia is the CosmOz network. This network consists of cosmic ray sensors operational at 2m above the ground where fast neutrons passing through the earth atmosphere are counted (Hawdon et al., 2014). These cosmic ray sensors are currently installed at nine locations throughout Australia. In fact, the probes count the neutrons in the soil and air above the soil. Soil water content controls primarily the counting of the neutrons as hydrogen atoms have a moderating effect on the fast neutron intensity. There is even an inverse correlation between neutron intensity and water content (Zreda et al., 2008). Neutron count is converted to soil moisture using a calibration function (Desilets et al., 2010). Data from one CosmOz station is used in this study.

One should keep in mind that Oznet and Ozflux measure soil moisture in the topsoil (0-10cm), whereas CosmOz has a varying soil moisture depth related to the wetness of the soil (Jackson et al., 2012). Moreover, Oznet and Ozflux are point measurements, while CosmOz has a spatial range of 300m around the probe (Hawdon et al., 2014). Contrast exists between sampling frequencies of the different soil moisture measurements. Oznet has a temporal resolution of 20min, Ozflux 30min and CosmOz hourly. Lastly, these measurement techniques are all accompanied by different error sources and uncertainties, which inhibits absolute agreement between measurements of these different sources (Brocca et al., 2009).

3 Results

For the evaluation of Tier-1 Earth2Observe output for the selected models, climate zones are grouped together based on the first letter for each climate zone and on the colors in Table 1. Am, As and Aw represent tropical climate zones. BWh, BSk, BSh represent arid climates. Cfa, Cfb, Csa, Csb and Cwa are corresponds to temperate climate zones. The national AWRA-L model serves as benchmark for the evaluation of the large-scale hydrological models. This model is originally at 0.05° spatial resolution. However, Tier-1 Earth2Observe simulations for runoff, actual evapotranspiration and soil moisture are all at 0.5° spatial resolution. Therefore, for direct comparison with the large-scale hydrological models, AWRA-L has been upscaled to 0.5° resolution. Also, it is important to note that AWRA-L is a calibrated model. For the classification of KGE values, Table 5 is used.

Tab. 5: Classification of KGE values for model evaluation

Classification	KGE score
Very good	$KGE > 0.7$
Good	$0.5 < KGE < 0.7$
Satisfying	$0.2 < KGE < 0.5$
Unsatisfying	$KGE < 0.2$

3.1 Tier-1 EartH2Observe model evaluation

3.1.1 Runoff

Runoff fields at 0.5° spatial resolution from the Tier-1 Earth2Observe project have been compared with streamflow observations. In general, monthly runoff fields perform better than daily when compared to the observations, see Table 6.

Tab. 6: Average runoff KGE values for both time resolutions listed for several climate zones in Australia. Comparisons have been made between observations and the large-scale hydrological models, the AWRA-L model and the multi-model ensemble median.

Climate-zone	W3RA		PCR-GLOBWB		HTESSEL		ORCHIDEE		AWRA-L		Median	
	monthly	daily	monthly	daily	monthly	daily	monthly	daily	monthly	daily	monthly	daily
Am	-0.03	-0.13	-0.05	-0.20	0.17	0.16	-0.63	-0.45	-0.52	-0.46	0.41	0.20
As	0.04	0.01	0.13	-0.01	0.21	0.25	-0.65	-0.51	0.17	0.14	0.46	0.33
Aw	0.22	-0.05	-0.09	-0.21	0.42	0.25	-0.69	-0.45	-0.57	-0.48	0.24	0.09
BWh	-0.55	-0.60	-1.36	-1.47	-0.20	-0.34	-14.81	-15.36	-14.20	-13.84	-0.33	-0.46
BSh	0.10	-0.20	-0.83	-0.96	0.14	-0.10	-10.29	-10.64	-5.66	-5.50	-0.05	-0.20
BSk	-9.52	-9.50	-10.57	-10.42	-3.43	-3.71	-391.97	-394.00	-373.38	-369.81	-3.54	-3.68
Cfa	-0.61	-0.85	-1.63	-1.94	-0.21	-0.77	-5.41	-6.20	-3.32	-3.36	-0.29	-0.53
Cfb	-0.50	-0.69	-1.63	-1.82	-0.12	-0.45	-4.90	-5.12	-12.05	-11.66	-0.39	-0.56
Csa	-7.69	-8.69	-6.87	-7.01	-1.24	-1.84	-15.37	-21.54	-12.79	-12.83	-2.35	-2.93
Csb	-2.91	-3.09	-4.38	-4.49	-0.38	-0.57	-5.65	-7.69	-19.43	-19.10	-0.94	-1.13
Cwa	0.18	0.00	0.03	-0.17	0.53	0.28	-2.21	-2.11	-0.17	-0.34	0.30	0.12

In order to make statements for each main climate group (i.e. tropical, arid and temperate), the average monthly KGE for each main climate group is calculated by taking the average of monthly KGE values for climate-zones belonging to that climate group:

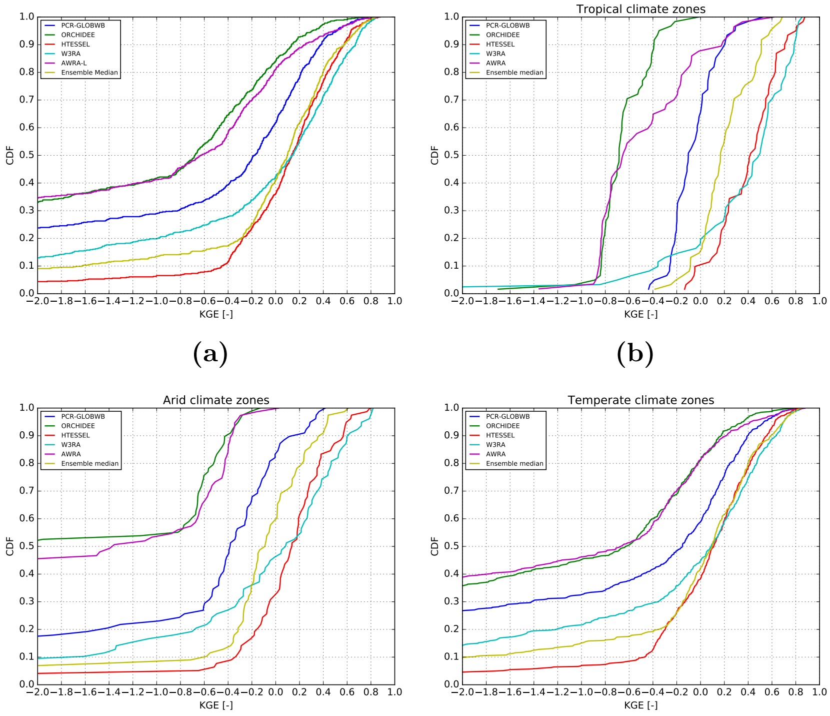
$$\overline{KGE}^{X}_{model} = \frac{1}{n}\sum_{i=1}^{n} KGE_{Xi}, \tag{7}$$

where X corresponds with the main climate group and i with the smaller climate-zones belonging to that particular climate group. For example, the KGE values for Aw, Am and As are summed and divided by three to obtain the average KGE value corresponding with the tropical climate zones.

For the large-scale hydrological models, only HTESSEL performs satisfying 5 in the tropics ($\overline{KGE}^{tropical}_{HTESSEL} = 0.27$). For the arid and temperate climate group, all large-scale hydrological models perform unsatisfying ($\overline{KGE} < 0.2$). However, HTESSEL performs good for the Cwa climate zone ($KGE^{Cwa}_{HTESSEL} = 0.53$) (Table 6).

Taking the ensemble median of the large-scale hydrological models leads to better monthly runoff estimates in the tropical areas compared to each individual model($\overline{KGE}^{tropical}_{Ensemble-median} = 0.37$ (Table 6). For the other climate zones, the ensemble median performs slightly worse than the best individual model (HTESSEL). Still, average performances for the arid and temperate climate zones (except Cwa for HTESSEL and ensemble median) are mainly unsatisfying (Table. 5) for all large-scale hydrological models, the ensemble median and the AWRA-L model. For the computation of average KGE values per climate zone, catchment area (”support”) is neglected. In this research, performances for small catchments are equally important as performances for larger catchments.

Very bad performances have severe influences on the average value for KGE. Therefore, its interesting to explore the distribution of performances for each large-scale hydrological model, the AWRA-L model and the ensemble median. Therefore, cumulative density plots (cdf's) have been made for the total monthly KGE distribution with all climate zones included (Fig. 4a) and for each climate zone separately (Fig. 4b-d). These cumulative density plots have been made from performances obtained from monthly streamflow simulations. These plots have only been constructed for monthly runoff fields as literature demonstrated that large-scale hydrological models generally perform better for monthly time resolution than for daily when compared with observations (Table 6). As such, large-scale hydrological models are better able to describe the hydrological processes occurring in a catchment for monthly simulations than for daily simulations. Taking all performances for all climate zones into account, results in 45% unsatisfactory performances (Table 4) for HTESSEL and W3RA (Fig. 4a). For the ensemble median, around 40% of the performances are unsatisfying. For PCR-GLOBWB, ORCHIDEE and AWRA-L, this value is 75%, 90% and 85% respectively. Another important aspect is that 10% of the performances are classified (Table 4) as very good for W3RA. This value equals 5% for both HTESSEL and the ensemble median, <2% for both PCR-GLOBWB and AWRA-L and <1% for ORCHIDEE.



(a)

(b)

(c)

(d)

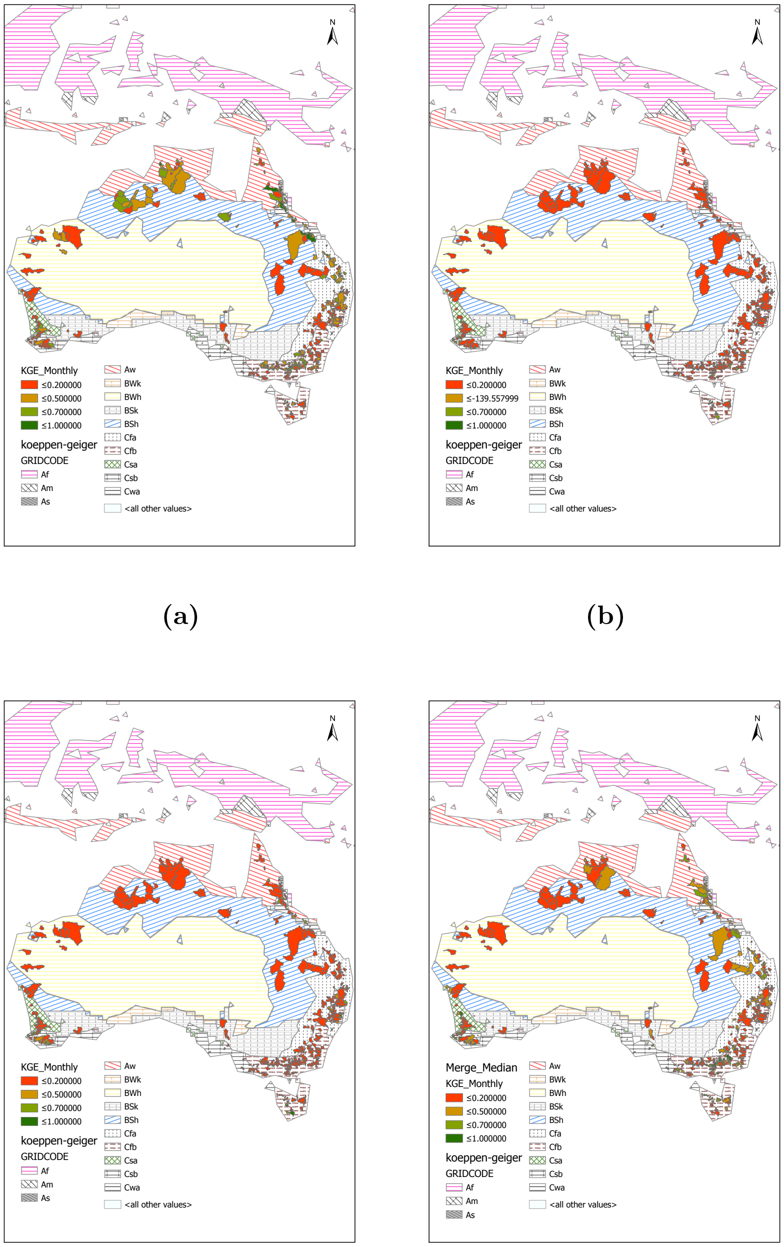
Fig. 4: Cumulative density plots for all climate zones together (a), tropical (b), arid (c) and temperate (d) based on monthly runoff fields.

Tropical climate zone For catchments from the tropical areas, W3RA has the biggest amount of high KGE scores for this climatic zone, with 25% very good performances (Fig. 4b). For HTESSEL, 10% of the performances are very good. The other large-scale hydrological models, AWRA-L and the ensemble median have no very good performances for this climate zone (Fig. 4b). Furthermore, the ensemble median performs for 50% of the simulations unsatisfactory, while 30% of HTESSEL and W3RA simulations scores unsatisfying. For the tropical climate zones, all performances are unsatisfying for ORCHIDEE (Fig. 4b).

Arid climate zone Compared with the tropical climate zone, the large-scale hydrological models have less very good performances for streamfow simulations in the arid climate zone (Fig. 4c). For W3RA, 5-10% of the simulations for the evaluated catchments perform very good. This value is 0-5% for both HTESSEL and the ensemble median. AWRA-L, PCR-GLOBWB and ORCHIDEE have no very good performances in the arid climate zone at all (Fig. 4c). For both W3RA and HTESSEL, 50% of the performances in arid climates are unsatisfying. PCR-GLOBWB performs unsatisfying for 90% of the arid climate catchments. For the ensemble median, 80% of the evaluated catchments obtain an unsatisfying score. AWRA-L and ORCHIDEE, have only unsatisfying score for catchments situated in the arid climate zone (Fig. 4c).

Temperate climate zone For this climate zone, very good performances are rare among all large-scale hydrological models with 5% for W3RA and the ensemble median and <3% for PCR-GLOBWB, AWRA-L, ORCHIDEE and HTESSEL (Fig. 4d). For HTESSEL, W3RA and the ensemble median, 60% of their simulations performed unsatisfactory (Fig. 4d). PCR-GLOBWB simulations resulted in 75% unsatisfying performances and both AWRA-L and ORCHIDEE 90%.

Performance maps have been made for the best large-scale hydrological model, the worst large-scale hydrological model, AWRA-L (benchmark) and the ensemble median in order to provide the reader a visualization of the performance distribution across Australia (Fig. 5a-d). The best model is here based on having the most satisfying ($KGE > 0.2$) and the most very good ($KGE > 0.7$) performances among all climate zones (fig. 4a). Based on the above mentioned definition, HTESSEL is the best model and ORCHIDEE the worst. The reader is referred to the Appendix for the monthly performance maps of PCR-GLOBWB and HTESSEL and the daily performance maps for all large-scale hydrological models (Fig. 15).



(a) **(b)**

(c) **(d)**

Fig. 5: Performance maps based on KGE for monthly streamflow simulations for HTESSEL (a), ORCHIDEE (b), AWRA-L (c) and the multi-model ensemble median (d).

This research found that all large scale models perform as good or even better than AWRA-L after comparing runoff simulations with observations. However, this may be due to the upscaling of the AWRA-L model, which causes loss of information as modeled output is averaged (Sablok & Haziz, 2008). Especially the variability ratio from the Kling-Gupta Efficiency index could be negatively affected by this aggregation process. Also, the very bad performances for the arid climate zones may due to poor respresentation of hydrological processes for these areas. One possible explanation for the bad performances in the arid climate zone lies in the models poor representation of partitioning the net precipitation into infiltration and runoff. Especially in semi-arid/arid conditions, runoff generated by infiltration excess is dominant and this process is poorly represented by the majority of the models except for W3RA and AWRA-L (Table 2). In addition, Arnell (2000) found that the runoff response in arid/semi-arid environment is very sensitive to alterations in precipitation. As a result, uncertainties from precipitation fields are propagated to larger uncertainties in simulations for runoff volumes (Güntner & Bronstert, 2003).

3.1.2 Actual evapotranspiration

The Tier-1 actual evapotranspiration fields from the four selected large-scale hydrological models have been evaluated for performance. The selected large-scale hydrological models are all at 0.5° resolution. To put the results into perspective, performances for the large-scale hydrological models will be compared with AWRA-L. For direct comparison of performances between the large-scale hydrological models and AWRA-L, AWRA-L actual evapotranspiration simulations have been upscaled (Sect. 2.3.1) from 0.05° to 0.5° resolution. Furthermore, note that AWRA-L is a calibrated model.

For actual evapotranspiration simulations, monthly time resolution simulations are generally more reliable than daily ones (Table 7).

Tab. 7: Average evapotranspiration KGE values classified per climate zone in Australia for daily and monthly time resolutions for all global models, the AWRA-L model and the median of the global models.

	W3RA		PCR-GLOBWB		HTESSEL		ORCHIDEE		Median		AWRA-L	
Climate-Zone	monthly	daily	monthly	daily	monthly	daily	monthly	daily	monthly	daily	monthly	daily
Aw	0.57	0.45	0.37	0.22	0.76	0.59	0.71	0.45	-0.1	0.17	0.54	0.38
BSh	0.53	0.29	0.44	0.29	0.32	-0.079	0.4	-0.27	0.14	0.13	0.53	0.48
BSk	0.26	0.19	0.77	0.62	0.61	-0.41	-1.25	-1.15	0.48	0.29	0.021	-0.0038
BWh	-0.3	-0.26	0.33	0.27	-0.51	-0.49	0.018	-0.28	0.58	0.3	-0.4	-0.41
Cfa	0.3	0.23	0.17	0.18	0.44	0.39	0.4	0.19	0.5	0.2668	0.014	0.073
Cfb	0.44	0.24	0.57	0.32	0.5	0.34	0.39	0.18	0.24	0.09	0.61	0.52

Like for runoff evaluation, we take the average of the KGE values based on monthly simulations for each climate zone based on main climate group (tropical, arid, temperate)(Table. 1) by using Eq. 7.

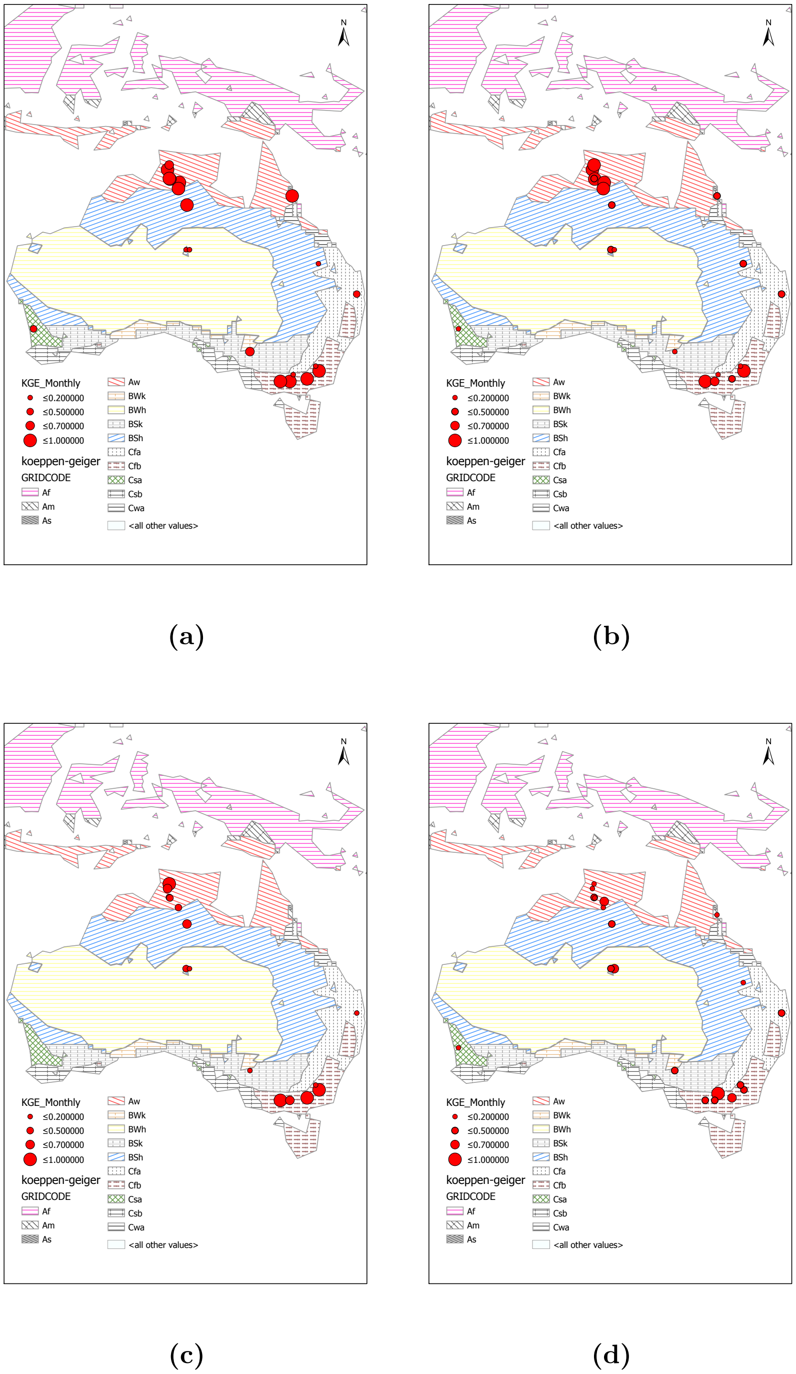
The LSMs have very good (Table 5) performances in the tropical regions ($\overline{KGE}^{Tropical}_{ORCHIDEE} = 0.71$ and $\overline{KGE}^{Tropical}_{HTESSEL} = 0.76$). However, for the arid climate zones, both ORCHIDEE and HTESSEL (LSMs) perform unsatisfactory (Table 5)($\overline{KGE}^{Arid}_{HTESSEL} = 0.14$ and $\overline{KGE}^{Arid}_{ORCHIDEE} = -0.28$) Contrary, PCR-GLOBWB (GHM) performs good (Table 5) in the arid climate zones ($\overline{KGE}^{Arid}_{PCR-GLOBWB} = 0.51$). W3RA has on average good scores for the tropical areas , but lower compared to both ORCHIDEE and HTESSEL($\overline{KGE}^{Tropical}_{W3RA} = 0.54$). All large-scale hydrological models perform on average satisfying (Table. 5) for the temperate climate zone ($\overline{KGE}^{Temperate}_{W3RA} = 0.37$; $\overline{KGE}^{Temperate}_{PCR-GLOBWB} = 0.37$; $\overline{KGE}^{Temperate}_{HTESSEL} = 0.47$; $\overline{KGE}^{Temperate}_{ORCHIDEE} = 0.40$)(Table. 7).

Compared with the national AWRA-L model, performances are comparable or even better ($\overline{KGE}^{Tropics}_{AWRA-L} = 0.54$; $\overline{KGE}^{Arid}_{AWRA-L} = 0.05$; $\overline{KGE}^{Temperate}_{AWRA-L} = 0.21$). However, as mentioned earlier (Sect. 2.3.1), upscaling induces loss in information (e.g. variability) and this relates to lower KGE values. The ensemble median has satisfying performances in the arid and temperate climate zones ($\overline{KGE}^{Arid}_{Ensemble-median} = 0.40$ and $\overline{KGE}^{Temperate}_{Ensemble-median} = 0.37$), but performs worse for the tropics than all other simulations (Table 7).

Fig. 6a-d visualizes the distribution of performances for the on average best performing model (HTESSEL), the worst performing (ORCHIDEE), AWRA-L and the multi-model ensemble median to give a general idea of how the performances are spatially distributed across Australia. These figures clearly show that both ORCHIDEE and HTESSEL perform better in the tropics

(Fig. 6). AWRA-L has its highest scores in the temperate climates and the ensemble median performs best in the arid climate zones (Fig. 6). The remaining monthly performance maps for the other large-scale hydrological models and performance maps for daily resolution are presented in the Appendix (Fig. 16). Also, daily time series figures for actual evapotranspiration for the large-scale models, AWRA-L, the ensemble median and the observations are made for every observation site separately and are included in the Appendix (Fig. 16, Fig. 17).

Bad performances in arid /semi-arid climate zones may be explained by difficulties in partitioning of net precipitation into infiltration and overland flow in those climate zones. This could result in underestimation of infiltration excess overland flow and overestimation of actual evapotranspiration. A possible explanation for the higher scores for the ensemble median could be that the multi-model ensemble median reduces model structure uncertainty (Ajami et al., 2007).



(a) **(b)**

(c) **(d)**

Fig. 6: Performance maps for actual evapotranspiration simulations from the Tier-1 Earth2Observe based on monthly KGE values for HTESSEL (a), ORCHIDEE (b), AWRA-L (c) and the multi-model ensemble median (d)

3.1.3 Soil Moisture - Bias

The Tier-1 EartH2Observe soil moisture estimates from the selected large-scale hydrological models have been evaluated for performance across Australia. The selected models from the Tier-1 Earth2Observe dataset for evaluation are all at 0.5° spatial resolution. Again, AWRA-L serves as benchmark for the performances of the large-scale hydrological models. This model is originally at 0.05°, but is upscaled (Sect. 2.3.1) to 0.5° spatial resolution for direct comparison with the large-scale hydrological models. In addition, AWRA-L is a calibrated model. For the evaluation of soil moisture, the bias term is excluded from the original KGE equation (Eq. 5) to overcome the differences in measurement/modeled depths by the measurements and the large-scale hydrological models (Sect. 2.4.3).

In general, monthly estimates for soil moisture give higher KGE scores for all model types for most climate zones in Australia (Table 8). This is an already known phenomenon in literature (Spruill et al., 2000).

Tab. 8: Average soil moisture KGE values for daily and monthly time resolutions is calculated for each climate zone. For soil moisture evaluation, the bias term is omitted from the origianl KGE equation (Eq. 5). Evaluation has been done for the large-scale hydrological models, the AWRA-L model and the multi-model ensemble median. Due to data availability, AWRA-L has less evaluated sites. As a result, Csa is not covered by AWRA-L.

	W3RA		PCR-GLOBWB		HTESSEL		ORCHIDEE		Median		AWRA-L	
Climate-Zone	monthly	daily	monthly	daily	monthly	daily	monthly	daily	monthly	daily	monthly	daily
Aw	0.42	0.36	-0.23	-0.29	-0.95	-0.64	-0.091	0.037	-0.65	-0.6	-0.047	-0.053
BSh	0.18	0.27	0.61	0.53	0.6	0.34	0.78	0.65	0.61	0.54	0.0051	-0.015
BSk	0.22	0.18	0.33	0.3	0.61	0.52	0.52	0.46	0.73	0.59	-0.0048	-0.03
Cfa	0.17	0.15	0.53	0.42	0.74	0.66	0.62	0.54	0.72	0.63	-0.017	-0.08
Cfb	-0.18	0.17	0.46	0.34	0.38	0.31	0.53	0.4	0.58	0.37	-0.029	-0.1
Csa	0.12	0.54	-0.21	-1.01	0.037	-1.13	0.18	0.63	0.13	-0.065	X	X

For the next analysis, KGE values from Table 8 are averaged for each main climate group (Eq. 7) (Table 1. According to Table 4, W3RA has satisfying scores in the tropical and arid climate zones ($\overline{KGE}_{W3RA}^{tropical} = 0.42$ and $\overline{KGE}_{W3RA}^{arid} = 0.20$), while PCR-GLOBWB has good scores in the arid and satisfying scores for the temperate climate zones ($\overline{KGE}_{PCR-GLOBWB}^{arid} = 0.47$; $\overline{KGE}_{PCR-GLOBWB}^{temperate} = 0.26$). However, PCR-GLOBWB performs unsatisfying in the tropical climate zone ($\overline{KGE}_{PCR-GLOBWB}^{tropical} = -0.23$). HTESSEL performs in line with PCR-GLOBWB with good scores in the arid climate, satisfying scores in the temperate climate zones and unsatisfying scores for the tropical climate zone ($\overline{KGE}^{arid}_{HTESSEL} = 0.61$; $\overline{KGE}^{temperate}_{HTESSEL} = 0.38$; $\overline{KGE}^{tropical}_{HTESSEL} = -0.95$). Like PCR-GLOBWB and HTESSEL, ORCHIDEE has good scores in the arid climate zones, satisfying performances in the temperate climate regions, but unsatisfying performances in the tropical climate regimes ($\overline{KGE}^{arid}_{ORCHIDEE} = 0.65$; $\overline{KGE}^{temperate}_{ORCHIDEE} = 0.44$; $\overline{KGE}^{tropical}_{ORCHIDEE} = -0.091$) (Table 8).

As mentioned in the general model inter-comparison framework (Sect. 2.3), performances of the large-scale uncalibrated model runs are compared with the calibrated AWRA-L model. However, AWRA-L performs unsatisfying in all climate zones ($\overline{KGE}^{arid,tropical,temperate}_{AWRA-L} < 0.2$) (Table 8). Moreover the spread of performances between the different climate zones is very low. The ensemble median has unsatisfying performances in the tropics ($\overline{KGE}^{tropical}_{Ensemble-median} = -0.65$), while performances are good and satisfying in the arid and the temperate climate zones respectively ($\overline{KGE}^{arid}_{Ensemble-median} = 0.67$; $\overline{KGE}^{temperate}_{Ensemble-median} = 0.48$). For the arid and temperate climate-zones, the ensemble median performs as good or even better than each individual model. However, W3RA and ORCHIDEE perform better for the tropical climate regions than the ensemble median (Table 8).

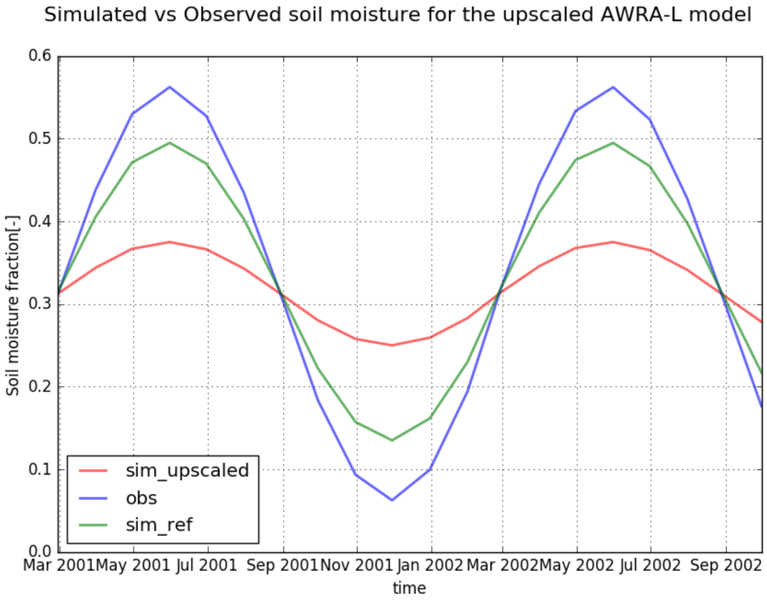


Fig. 7: Example of differences in variability between soil moisture observations and AWRA-L simulations due to upscaling

Fig. 8a-d) illustrate the distribution of model performances for soil moisture estimation throughout Australia. As with actual evapotranspiration evaluation, performance distributions for the best and the worst large-scale hydrological model, the AWRA-L model and the ensemble median are presented here. KGE values in these Fig. 8a-d are based on monthly time resolution analysis. The monthly time resolution performance maps for the other large-scale hydrological models and all daily time resolution performance maps for the large-scale hydrological models are given in the Appendix (Fig. 18, Fig. 18).

According to Table 8, AWRA-L performs unsatisfying for each climate zone. As literature demonstrated (Sablok & Aziz, 2008), upscaling model outputs leads to loss of information. Due to the variability and non-linearity of the soil moisture state and associated physics, this effect could be severe for the simulated AWRA-L variability of soil moisture fields. Fig. 7 illustrates the possible effect of upscaling on the time evolution of soil moisture. Pearsons correlation values are generally very good for the AWRA-L (Table 16), which further underpins above mentioned statement.

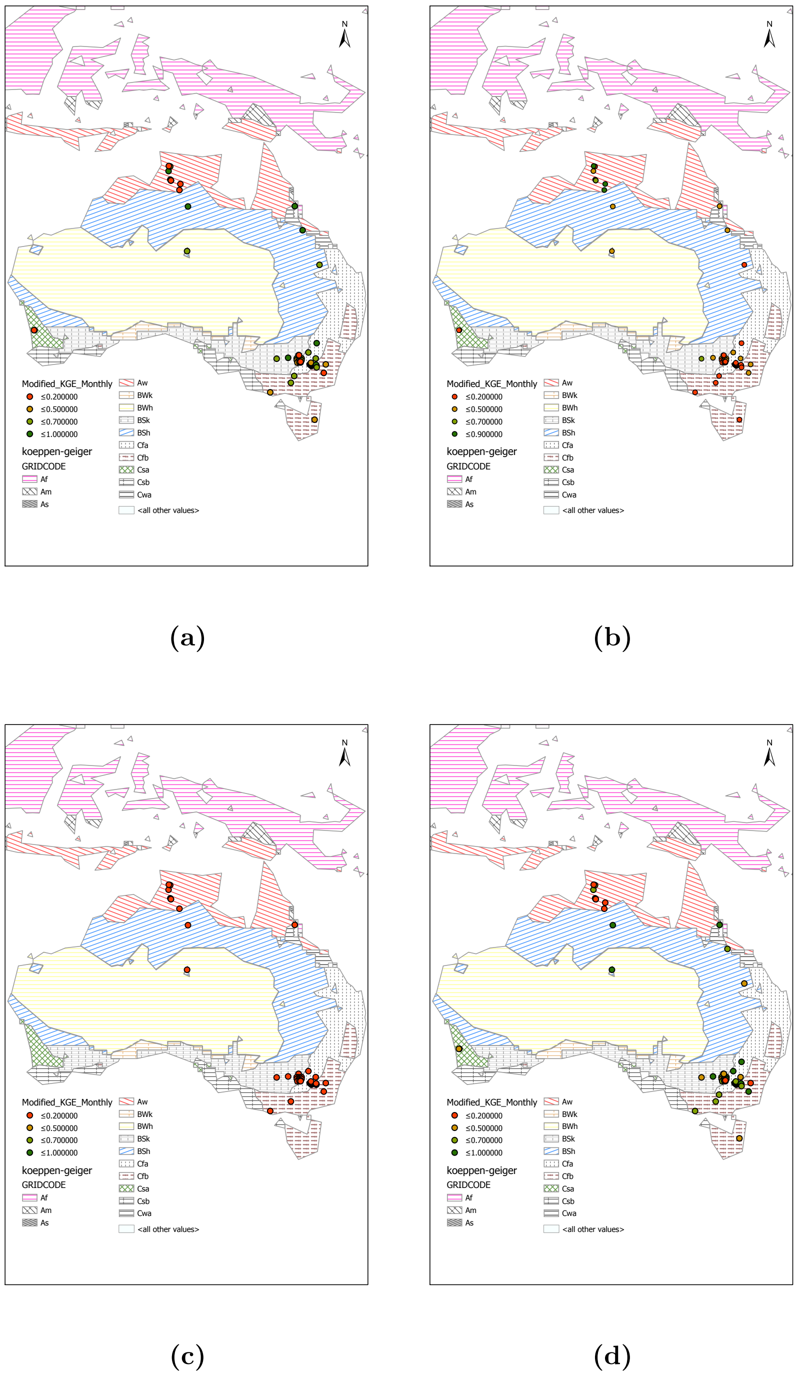


Fig. 8: Soil moisture monthly modified KGE values for ORCHIDEE (a), W3RA (b), AWRA-L (c) and the ensemble median (d) are presented.

3.2 Summary Earth2Observe Tier-1 evaluation

Table 9 summarizes the best and worst performing large-scale hydrological for each main climate group (tropical, arid, temperate) for each evaluated variable. Also, AWRA-L and the ensemble median are included in this comparison.

Tab. 9: Worst and best large-scale hydrological model performances for the Tier-1 Earth2Observe dataset for runoff, soil moisture and actual evapotranspiration. KGE is used as performance metric and for soil moisture evaluation, the bias between simulations and observations is excluded from the KGE equation (Eq. 5).

	Best model			**Worst model**		
Variable	Tropical	Arid	Temperate	Tropical	Arid	Temperate
Runoff	W3RA	W3RA/HTESSEL	W3RA	ORCHIDEE	ORCHIDEE	AWRA-L
Actual evapotranspiration	HTESSEL/ORCHIDEE	PCRGLOB-WB	HTESSEL	Ensemble Median	ORCHIDEE	AWRA-L
Soil moisture	W3RA	Ensemble median	Ensemble median	HTESSEL	AWRA-L	W3RA

So, for the evaluation of runoff fields, W3RA performed best in the tropical climate zones. ORCHIDEE obtained the worst scores for this climate zone. W3RA and and HTESSEL simulations resulted in the best performances in the arid climate zones, while ORCHIDEE performed worst in this climate zone. For the temperate climate zone, W3RA performed best and AWRA-L worst (Table 9). For actual evaporation, the best performances for the tropical climate zones are obtained by HTESSEL and ORCHIDEE. The ensemble median obtained the lowest KGE values for this climate zone. For the arid climate zone, PCR-GLOBWB performed best, whereas ORCHIDEE simulations resulted in the lowest scores for this climate zone. HTESSEL performed best for the temperate climate zones, whereas the lowest scores for the temperate climate zones are obtained by AWRA-L (Table 9). After evaluating soil moisture fields, best performances are obtained W3RA for the tropical climate zones. HTESSEL performed worst for this climate zone. The ensemble median performed best for both the arid and the temperate climate zones, whereas AWRA-L performed worst in the arid climate zone and W3RA in the temperate climate zones (Table 9).

3.3 Calibration of PCR-GLOBWB

In section 3.1.1 we evaluated four large-scale hydrological models. The differences found between simulations by the evaluated large-scale hydrological models are related to model structure uncertainty, because input forcing and observational data are equal for all large-scale hydrological models from the Tier-1 Earth2Observe project.

Large-scale hydrological models have a large number of parameters, which could contain uncertainty. In order to decrease the parameter uncertainty and increase the reliability of model predictions, calibration and validation of the large-scale hydrological model is an essential step to improve the ability to represent the real world processes more accurately. This section is dedicated to the calibration/validation of PCR-GLOBWB.

Four catchments have been selected based on several criteria, which are mentioned in section 2.3.2. For the selection of catchments for PCR-GLOBWB calibration, a minimum requirement of $KGE > -0.5$ is set for the Tier-1 uncalibrated PCR-GLOBWB run to ensure that the model structure is representing the hydrological processes sufficiently. We argue that very low KGE for the values for the uncalibrated PCR-GLOBWB run are due to either a poor forcing dataset or a poor model structure instead of bad parameterization of the hydrological model (Sect. 2.3.2) Catchment characteristics for the selected catchments are listed in Fig. 10.

Tab. 10: Catchment characteristics of the selected catchments used for calibration of the PCR-GLOBWB model

Name	Area (km^2)	Min elevation (m)	Max elevation (m)	Climate-zone	Mean P (mm)	Mean ET$_p$(mm)
Clarence River	16953	12	1560	Cfb/Cfa	1057	1408
De Grey River	53323	22	654	BWh	393	1769
Gregory River	11291	120	432	BSh	501	1753
Roper River	43476	12	439	Aw/Bsh	878	2111

The selected catchments for calibration are distributed over different climate zones in order to minimize the effect of parameterization of the model for processes related to a certain climate zone. As such, this decreases the effect of the "regionalization problem", where parameters sets for certain catchments may not be representative for other catchments. Criteria for the catchment selection are given in section 2.3.2. The catchments are located in tropical

(Aw), semi-arid (Bsh), arid (Bwh) and temperate climate zones (Cfb).

For the calibration of PCR-GLOBWB, monthly streamflow data from these four catchments in Australia are used. The monthly simulated and observed hydrographs of a subset of catchments are used to infer the parameters, which might positively influence the performance of PCR-GLOBWB simulations. These are called effective parameters. The monthly simulated vs. observed hydrographs are used for information regarding the selection of effective hydrological model parameters. These hydrographs are shown in Fig. 9. Based on these hydrographs, $StorCap$, K_{sat} and J are the selected effective parameters for the brute-force calibration procedure. Table 11 describes multiplication factors for the parameters used in the calibration process.

Tab. 11: Parameters used for calibration of PCR-GLOBWB model

Parameter	**Long name**	**Units**	**Multiplier values**
$StorCap$	Storage capacity	m^3	0.75, 1.00, 1.25
J	Base flow recession coefficient	s^{-1}	-0.5, 0, 0.5, 1.0
K_{sat}	Saturated conductivity	m/s	-0.25, 0, 0.25

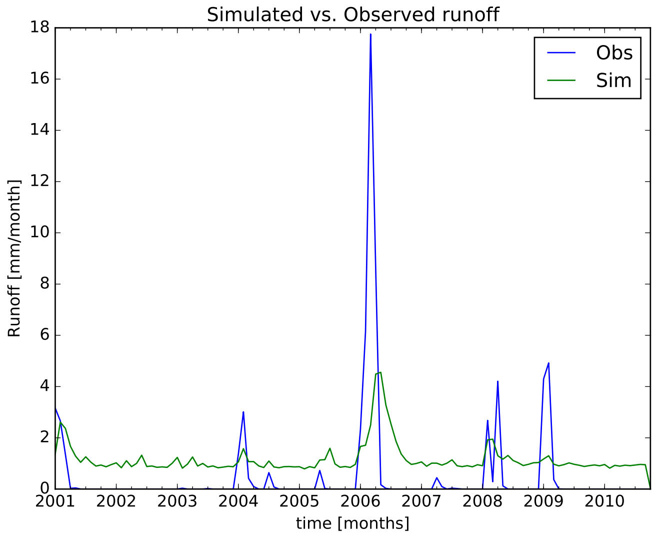
$StorCap$ is storage capacity $[m^3]$, K_{sat} is the saturated conductivity $[m/s]$ and J represents the recession coefficient $[s^{-1}]$. As mentioned earlier, we define for these parameter multipliers the min/max values and the increments. After, all possible parameter combinations are made within a script and these parameter sets are used as input for several PCR-GLOBWB runs. The KGE will be used as objective function as it tests the temporal agreement, variability and the absolute difference between simulations and observations (Eq. 5).

Table 12 provides all parameter combinations with their corresponding KGE values. Run016 is the reference run and this parameter set has been used for the EartH2Observe project. This reference scenario has been evaluated for performance in section 3.1.1 of this thesis. For the calculation of the average KGE value for each run among the four catchments, we decided to take KGE values above 0 into account only (Eq. 6). In our opinion, runs with very bad KGE values for catchments are related to bad model structure or poor forcing data for those areas rather than wrong parameterization.

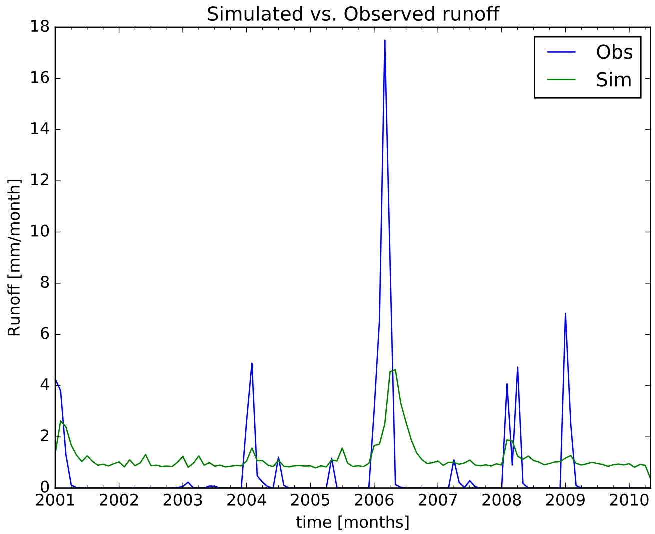
Therefore, no differentiation between bad and very bad performances for certain catchments in different runs has been made. To exclude this effect, Eq 6 is used, which is described in section 2.3.3 of this thesis. This research uses a global calibration procedure, in which one parameter set is sought for all climate zones (Sect. 2.3.2). Therefore, for each PCR-GLOBWB run, the average of the KGE value for the selected catchments will be calculated.

As seen in Table 12, run010 has the best performance among these catchments with a KGE value of 0.165. The average KGE value for the run with the default parameter set is 0.052. As mentioned in section 2.3.2, the range of behavioral sets is defined as the top 5% of average KGE performances for parameter combinations. Based on the average KGE column (Table 12), the top 5% equals 0.1487. As a result, the parameter combination of run000 is considered behavioral as well.

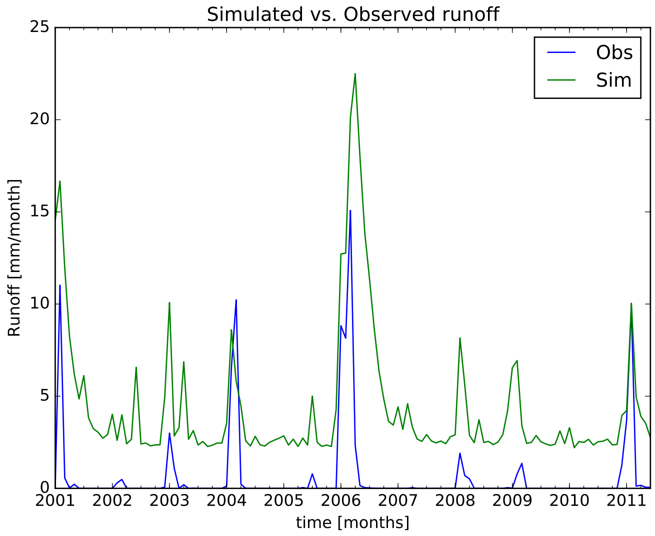
It is interesting to see how the change in parameter affects average KGE value among these catchments. Therefore, Fig 10 is made, where KGE values for all runs for each parameter are shown. A parameter is identifiable when the average of the objective function (KGE) for all runs changes as parameters increase or decrease. According to Fig 10, the PCR-GLOBWB model improves by a decrease in the prefactor value for Ksat. The recession coefficient prefactor, J, shows a large spread among all runs and is therefore non-identifiable. The prefactor for the storage capacity parameter (StorCap) is well identified as generally lower values lead to higher KGE values for all runs. However, as indicated in Table 12, run010 with prefactor value 1.0 for storage capacity has the highest KGE value. This is also visible in Fig 10. Therefore, the run with $f_j = 1.0$, $f_k = -0.25$ and $f_s = 1.0$ (run010) is chosen as the calibrated run based on average KGE for the four chosen catchments across Australia. The ranges of behavorial parameters is: $StorCap[0.75 - 1.00], K_{sat}[-0.25]$ and $StorCap[0.0]$ (Table 12). If we look at Table 12, the higher average KGE values are found for $f_k = -0.25, f_s = 0.75$. In fact, for these four selected catchments, there is a bias for PCR-GLOBWB model performance when using these parameter prefactors. This is also visible in Fig. 10. However the highest value for average KGE is generated by run010 and this run will be validated in the next section (Sect. 3.4.1). So, based on this global calibration strategy, PCR-GLOBWB needs an higher recession coefficient and a lower Ksat. Firstly, this means that after a storm event, the model needs to simulate a prolonged decline in streamflow. Secondly, a lower multiplier for K_{sat} means that, during and after a rainfall event, less infiltration due to less percolation occurs and this results in more generated runoff in the model (saturation excess overland flow).



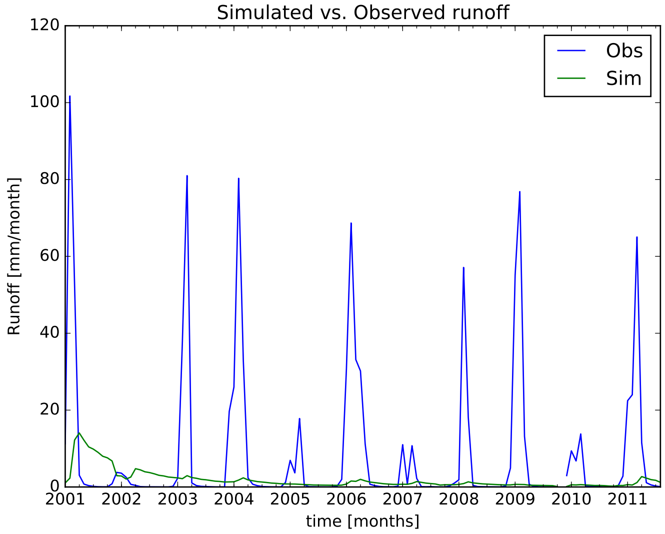
(a)



(b)



(c)



(d)

Fig. 9: In fig a-d monthly simulated vs. observed runoff for the 4 biggest catchments in Australia are shown.

Tab. 12: Parameter combinations for calibration of the PCR-GLOBWB model for catchments in Australia based on monthly streamflow simulations. The optimal parameter combination is indicated in green and the default parameter setting in blue.

Run number	Ksat	J	Storcap	average monthly KGE
run000	-0.25	-0.5	0.75	0.155
run001	-0.25	-0.5	1.0	0.103
run002	-0.25	-0.5	1.25	0
run003	-0.25	0	0.75	0.124
run004	-0.25	0	1.0	0.126
run005	-0.25	0	1.25	0
run006	-0.25	0.5	0.75	0.129
run007	-0.25	0.5	1.0	0.146
run008	-0.25	0.5	1.25	0
run009	-0.25	1.0	0.75	0.127
run010	-0.25	1.0	1.0	0.165
run011	-0.25	1.0	1.25	0
run012	0	-0.5	0.75	0.113
run013	0	-0.5	1.0	0.028
run014	0	-0.5	1.25	0
run015	0	0	0.75	0.099
run016	0	0	1.0	0.052
run017	0	0	1.25	0
run018	0	0.5	0.75	0.099
run019	0	0.5	1.0	0.078
run020	0	0.5	1.25	0
run021	0	1.0	0.75	0.104
run022	0	1.0	1.0	0.105
run023	0	1.0	1.25	0
run024	0.25	-0.5	0.75	0.098
run025	0.25	-0.5	1.0	0.0085
run026	0.25	-0.5	1.25	0
run027	0.25	0	0.75	0.106
run028	0.25	0	1.0	0.036
run029	0.25	0	1.25	0
run030	0.25	0.5	0.75	0.115
run031	0.25	0.5	1.0	0.058
run032	0.25	0.5	1.25	0
run033	0.25	1.0	0.75	0.137
run034	0.25	1.0	1.0	0.093
run035	0.25	1.0	1.25	0.003

In Table 12, KGE values are averaged over the different catchments and the parameter set with the maximum average value is chosen as the calibrated parameter set. Table 13 shows the KGE values for each run and for each catchment individually. Local optimal runs are indicated in bold and these runs have the highest KGE value among all runs for that certain catchment. Table 13 shows that run010, which is the calibrated run based on the global calibration procedure, is also a local optimal run for the Clarence river catchment, which is located in both the temperate and semi-arid climate zone (Table 3). By calibrating PCR-GLOBWB, minor improvements have been achieved for De Grey catchment, which is situated in the arid climate zone. Therefore, bad performances for this catchment are either related to a wrong selection of effective parameters for this climate-zone or caused by poor forcing data or model structure rather than a sub-optimal parameter combination for the selected effective parameters.

Tab. 13: KGE performances for monthly streamflow simulations for the reference run together with PCR-GLOBWB runs with different parameter settings. Local optimal runs are indicated in bold

Catchment names	**Roper river**	**De Grey river**	**Clarence river**	**Gregory river**
Run ID				
run016 (reference)	-0.308	-0.499	0.151	0.0584
run000	0.0267	-0.441	0.121	**0.471**
run001	-0.282	-0.595	0.0981	0.314
run002	-0.497	-0.700	-0.172	-0.0489
run003	-0.263	-0.344	0.137	0.357
run004	-0.200	-0.567	0.174	0.330
run005	-0.524	-0.701	-0.122	-0.0466
run006	-0.449	-0.253	0.126	0.390
run007	-0.158	-0.532	0.242	0.343
run008	-0.566	-0.702	-0.0709	-0.0417
run009	-0.474	-0.231	0.102	0.407
run010	-0.0754	-0.529	**0.283**	0.377
run011	-0.567	-0.701	-0.0385	-0.0299
run012	-0.105	-0.411	0.169	0.282
run013	-0.417	-0.534	0.0588	0.0304
run014	-0.498	-0.630	-0.198	-0.349
run015	-0.126	-0.311	0.174	0.224
run017	-0.535	-0.631	-0.126	-0.338
run018	-0.304	-0.234	0.141	0.256
run019	-0.219	-0.484	0.231	0.0800
run020	-0.557	-0.644	-0.0606	-0.339
run021	-0.370	**-0.214**	0.0878	0.314
run022	-0.0824	-0.507	0.277	0.142
run023	-0.519	-0.655	-0.0226	-0.340
run024	-0.344	-0.468	0.210	0.181
run025	-0.449	-0.548	0.0339	-0.114
run026	-0.493	-0.634	-0.180	-0.397
run027	-0.150	-0.400	0.194	0.230
run028	-0.360	-0.518	0.143	-0.0584
run029	-0.512	-0.633	-0.0973	-0.386
run030	0.0690	-0.331	0.116	0.274
run031	-0.256	-0.495	0.230	-0.0120
run032	-0.514	-0.646	-0.0268	-0.395
run033	**0.175**	-0.309	0.00859	0.365
run034	-0.0883	-0.506	0.280	0.0924
run035	-0.452	-0.652	0.0118	-0.378

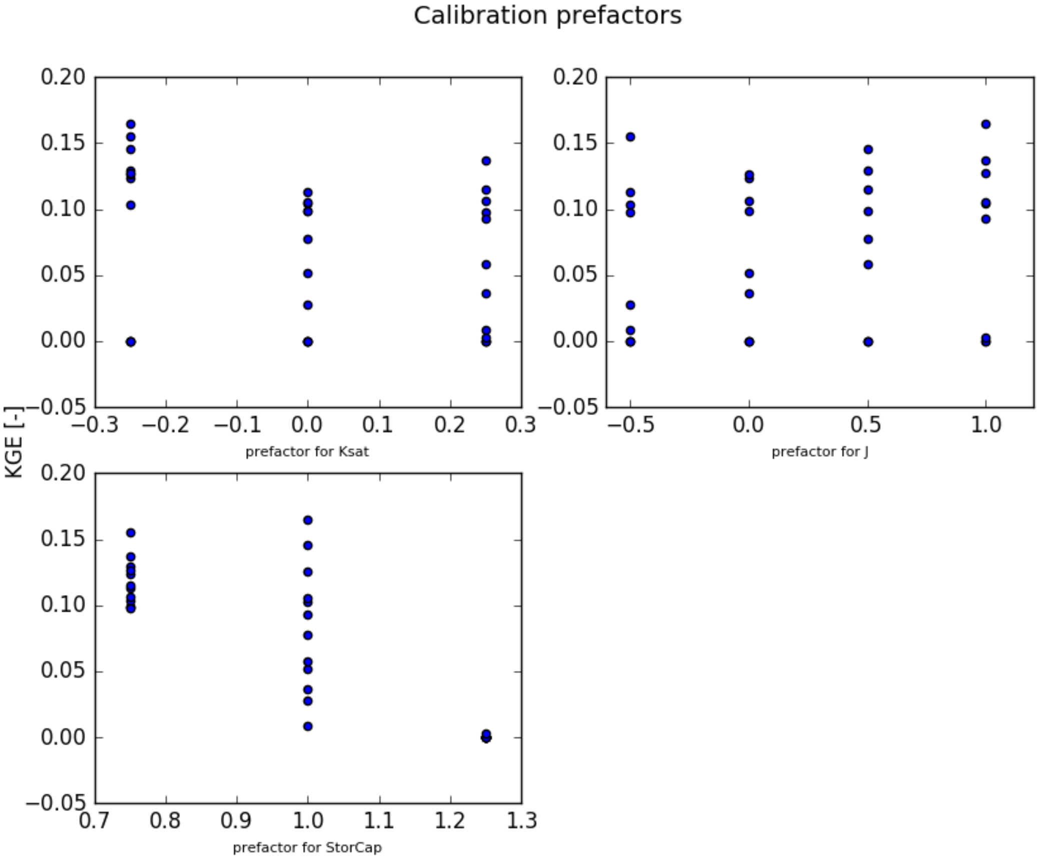
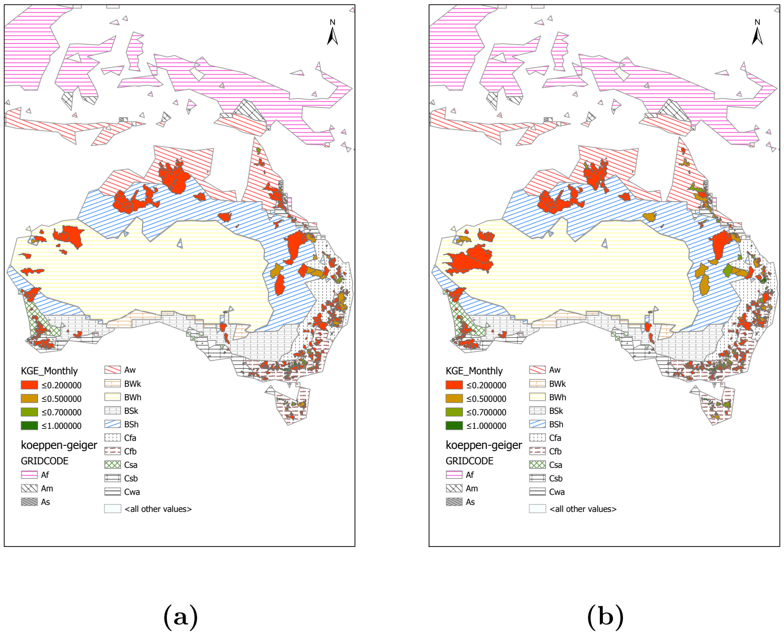


Fig. 10: KGE values for all combinations of calibration prefactors based on monthly runoff fields.

3.4 Validation

3.4.1 Runoff

After run010 has been chosen as the calibrated run, validation of the PCR-GLOBWB model with this parameter set is needed for all other catchments. To validate this calibration scenario, performances of the monthly simulated discharge time series of the PCR-GLOBWB model is evaluated for all other unimpaired catchments in Australia. Fig. 11 corresponds with KGE values for monthly time resolution for catchments in Australia using the reference run (Fig. 11a) and the calibrated run (Fig. 11b). Validation of the calibrated PCR-GLOBWB model is performed for all catchments except the ones used for calibration. If we compare Fig. 11b with the reference run as shown in Fig. 11a, it is easy to see major improvements for catchments in the tropical, subtropical and the arid climate zones.



(a) **(b)**

Fig. 11: Validation of monthly simulated runoff by comparison of KGE values between of the reference PCR-GLOBWB model (a) and the calibrated PCR-GLOBWB model (b) for catchments in Australia

In Fig. 12, a scatterplot for catchment KGE values between the uncalibrated PCR-GLOBWB model run and the calibrated PCR-GLOBWB model run is shown. This plot excludes the very bad performances ($KGE < 0$), as these performances are to a large extent due to poor representation of hydrological processes or unreliable forcing data. If points on this plot are close to the 1:1 line, no substantial improvements have been made by calibrating the model. However, points are predominantly located under the 1:1 line, which proves the success and importance calibrating PCR-GLOBWB (Fig. 12).

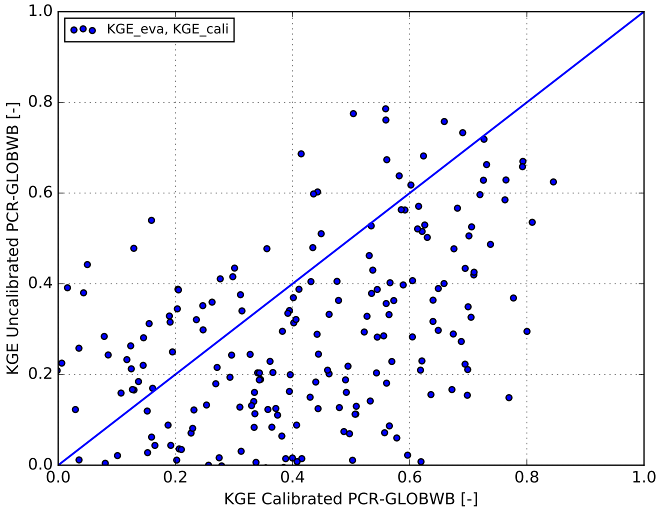


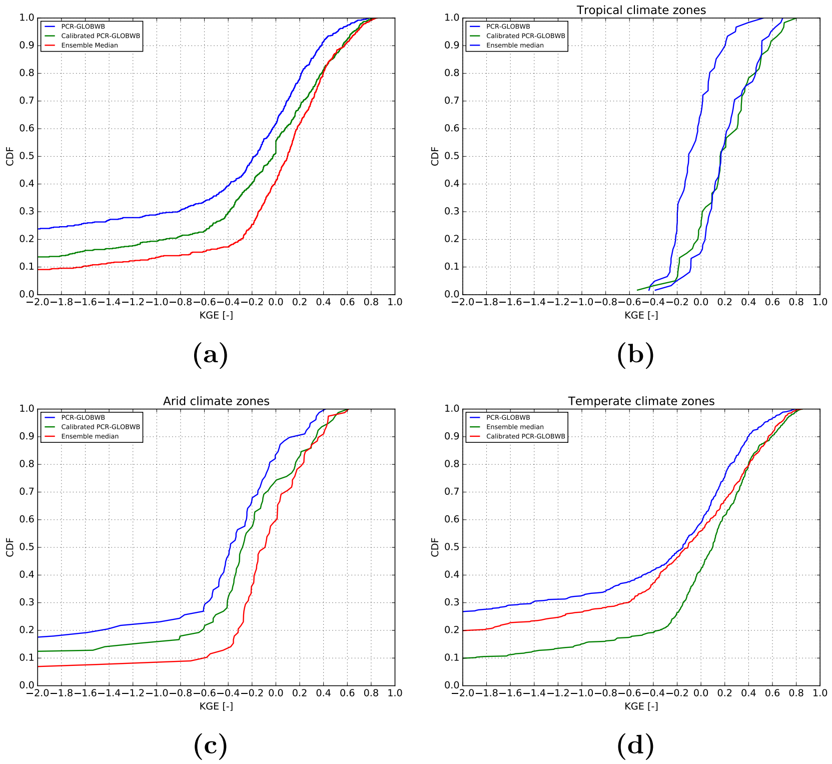
Fig. 12: Scatterplot of KGE value pairs between uncalibrated and calibrated PCR-GLOBWB model runs for catchments in Australia

Cumulative distribution functions have been made for performances from all catchments and for performances per climate zone (fig. 13a-d). In Fig. 13a, overall improvements have been achieved for streamflow simulations by calibrating the PCR-GLOBWB model. The calibrated PCR-GLOBWB run leads to a decrease of unsatisfying simulations compared to the default parameter setting. There are fewer really bad performing simulations compared to the reference uncalibrated run. Taking all climate zones into account, the reference run has 80% unsatisfying performances, while the calibrated run has 65% unsatisfying performances. Moreover, the calibrated PCR-GLOBWB model has a comparable amount of good ($0.5 < KGE < 0.7$) and very good ($KGE > 0.7$) performing simulations as the ensemble median.

Tropical For the tropical climate zones, the calibrated PCR-GLOBWB model performs unsatisfactory for 55% of its simulations compared to 90% for the default parameter setting (Fig. 13). The ensemble median has unsatisfying results for 55% of its simulations as well. Furthermore, none of the simulations from the uncalibrated PCR-GLOBWB model could be classified as good performing, while 15-20% of the calibrated PCR-GLOBWB simulations are good performing for this climate zone (Fig. 13b). The amount of good performing streamflow simulations for the calibrated PCR-GLOBWB run is even higher than the ensemble median (<10%).

Arid According to Fig 13c, the performances have improved after calibrating the PCR-GLOBWB model. 80-85% of the simulations performed unsatisifying compared to 90% for the reference run. There were no good performing simulations for the reference PR-GLOBWB run, whereas the calibrated run has approximately 5%. Still, for this climate zone, the ensemble median performs slightly better than the calibrated PCR-GLOBWB model.

Temperate For the temperate climate zone, performance improvements have been achieved by calibrating the PCR-GLOBWB model as well. According to Fig. 13d, the calibrated model has 10% less unsatisfying streamflow simulations for this climate zone (65% instead of 75%). Furthermore, the calibrated PCR-GLOBWB model has a 10% increase in good performing simulations compared to the uncalibrated PCR-GLOBWB model (15% instead of 5%). Compared to the ensemble median, the calibrated PCR-GLOBWB model has 5% more unsatisfying simulations. However, calibration of the PCR-GLOBWB had led to an equal amount of good performing simulations (fig 13d). This climate zone experienced a global improvement, despite its mixed uncalibrated performance run (Tier-1 Earth2Observe, Sect. 3.1.1).



(a) **(b)**

(c) **(d)**

Fig. 13: Cumulative density plots for all climate zones (a), tropical (b), arid (c) and temperate (d).

Improvements for the calibrated PCR-GLOBWB simulations for temperate climate zone catchments compared to the reference run seems logical as run010 is the local optimal run for the Clarence river catchment, which is situated in the temperate climate zone. Furthermore, improvements made for the tropical areas is explained by a higher KGE value for the Roper river catchment (Aw/BSh) for run010 compared to the reference run. Improvements in model performance after calibrating PCR-GLOBWB for the arid climates are less pronounced, which could be explained by the fact that run010 has equally bad performance for De Grey River catchment (BWh climate-zone) as the reference run. Fig. 11 visualizes the KGE values for both the uncalibrated and the calibrated PCR-GLOBWB model throughout Australia. This map complements the CDF curves (Fig. 13b-d) for the different climate zones, as improvements are made primarily in the tropical and the temperate climate zones. In Table 14, averages for each climate zone have been calculated. These averages are compared with the reference run (run016). As expected, the calibrated PCR-GLOBWB model performs better in climate zones for which the model is calibrated. As mentioned in the calibration subsection (Sect. 2.3.2), four sufficiently performing catchments ($KGE > -0.5$) have been chosen from tropical, sub-tropical, arid climate zones and temperate climate zones. Improvements have been made for all climate zones, which were used for calibration of PCR-GLOBWB (Aw, Cfb, BWh and Bsh, local improvement). Also, performances for other climate zones improved, which is a global improvement. However, many improvements have been made from "very bad" to "bad". So, one could assume that weak performances in those areas are due to poor representation of hydrological processes or unreliable forcing data.

Tab. 14: Reference vs. calibration scenario KGE monthly values for runoff across different climate zones in Australia.

Climate zone	Reference PCR-GLOBWB (run016)	Calibrated PCR-GLOBWB (run010)
Am	-0.05	0.23
As	0.13	0.31
Aw	-0.09	0.18
BWh	-1.36	-0.84
BSk	-10.57	-8.32
BSh	-0.83	-0.42
Cfa	-1.63	-0.68
Cfb	-1.63	-1.14
Csa	-6.87	-6.45
Csb	-4.38	-4.28
Cwa	0.03	0.24

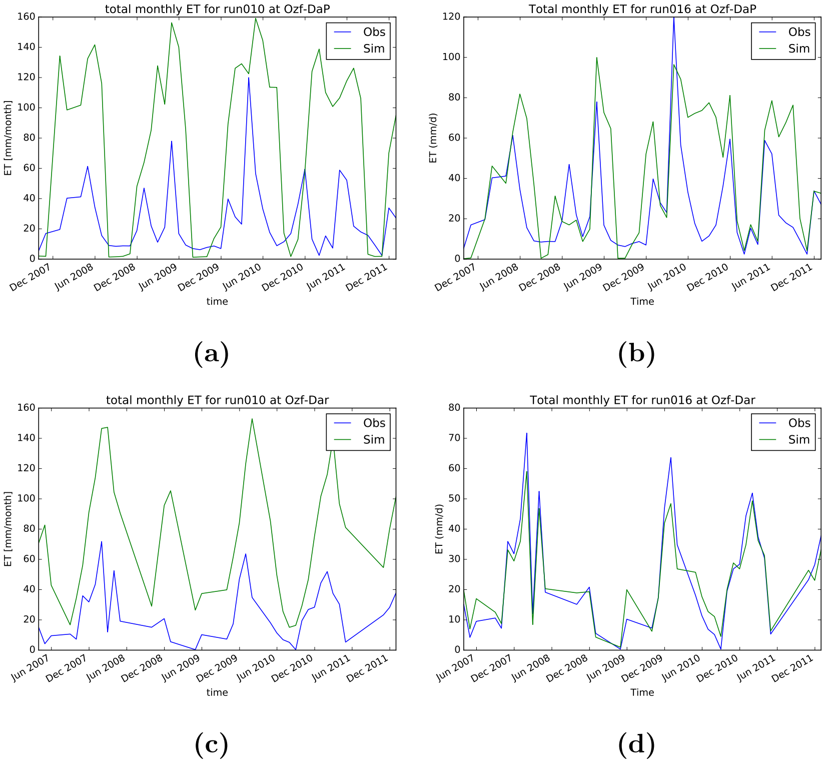
3.4.2 Actual evapotranspiration

Taking a look at the influence of PCR-GLOBWB calibration using streamflow observations (run010) from Australia on the performance of actual evapotranspiration simulations, it is possible to point out a decrease in performance (Table 15). The calibrated PCR-GLOBWB model still simulates the temporal evolution of actual evapotranspiration, but systematically overestimates the actual evapotranspiration quantitities. As a result, KGE values are lower for all sites. This could be partly explained by a lower prefactor for saturated conductivity, $f_k = -0.25$ (Fig. 10, Table 12). A lower prefactor for f_k leads to less percolation, which increases the soils ability to hold water, which increases actual evapotranspiration. However, storage capacity is decreased, but this has less effect on actual evapotranpsiration than the lower prefactor for saturated conductivity in run010. Fig. 14a-d shows the actual evapotranspiration time series for two Ozflux sites to illustrate the overestimation of actual evapotranspiration by run010, which is the calibrated PCR-GLOBWB for streamflow. Fig. 14a and Fig. 14c demonstrate that run010 is able to simulate the temporal dynamics of actual evapotranspiration for these Ozflux sites. However, run010 systematically overestimates actual evapotranspiration for these two sites. This leads to a bias between the observations and the actual evapotranspiration field for run010 and this translates into the lower KGE values than for the default parameter setting.

Concluding, the calibration of a hydrological model for a certain hydrological variable goes at the expense of the model performance for other hydrological variables. Therefore, before calibrating a hydrological model, it is important to keep the purpose of modeling in mind.

Tab. 15: Validation for actual evapotranspiration for the calibrated PCR-GLOBWB model (run010) reference model (run016) and another behavorial parameter set (run000)

Climate-Zone	Monthly correlation run010	Monthly KGE run010	Monthly correlation run016	Monthly KGE run016
Aw	-0.058	-1.47	0.54	0.42
Bsh	0.52	-1.03	0.56	0.54
Bsk	0.88	0.30	0.89	0.77
Aw	0.45	-1.53	0.59	0.16
Cfb	0.66	-1.48	0.97	0.78
Aw	0.32	-0.88	0.75	0.40
Aw	0.10	-1.068	0.78	0.45
Aw	0.32	-2.10	0.99	0.43
Cfb	0.53	-0.59	0.89	0.63
Cfa	0.59	-1.25	0.91	0.17
Bsh	0.62	-0.36	0.77	0.70
Bwh	0.61	-1.50	0.73	0.33
Cfb	0.62	-2.72	0.97	0.12
Cfb	0.41	-0.28	0.73	0.64
Cfb	0.79	0.31	0.97	0.40



(a) **(b)**

(c) **(d)**

Fig. 14: Actual evapotranspiration simulations for the calibrated PCR-GLOBWB at Ozfux site Daly River Pasture (a) and Ozflux site Daly river Uncleared (c) and PCR-GLOBWB run using default parameter setting for Ozfux site Daly river Pasture (b) and Daly river Uncleared (d). (b).

4 Discussion

The Earth2Observe project aimed to make a 30 year water resources reanalysis to insights in the full extent of water availability and the existing pressures t the regional/global scale. In fact, there is a need for large-scale hydrological modeling in order to make water resources assessments on a regional to global scale for to enhance water management, sustainable water use and in order to respond to natural hazards in the future. This thesis has evaluated four uncalibrated global hydrological models: PCR-GLOBWB (GHM), W3RA (GHM), HTESSEL (LSM) and ORCHIDEE (LSM) for performance. The large scales models are all at 0.5° spatial resolution. In this research, the national AWRA-L model is used as benchmark for the performances of the large-scale hydrological models. AWRA-L is a calibrated model and at 0.05° spatial resolution and is upscaled to 0.5° for direct comparison with the large-scale hydrological models. Performance evaluation has been done for runoff, actual evapotranspiration and soil moisture. This research used the Kling-Gupta Efficiency index as statistical metric to quantify the reliability of model simulations. For soil moisture, the bias is excluded from the original Kling-Gupta Efficiency equation as the large-scale hydrological models and the observations modeled/measured soil moisture over different depths.

One of the major findings were that monthly simulations by large-scale hydrological models are more reliable than daily simulations. This is observed for all three hydrological variables and confirmed in literature (Spruill et al., 2000). For runoff simulations, according to Mutenyo et al. (2013) and Spruill et al. (2000), monthly simulations are generally better than daily ones due to the inability of the large-scale hydrological models to capture peak flows.

A second important finding is that this research demonstrated that the ensemble median of the large-scale hydrological models performed best in tropical regions, whereas ORCHIDEE performed worst in the tropical climate zones. Both W3RA and HTESSEL performed best in the arid climate zone. ORCHIDEE performed worst in this climate-zone. Earlier research (Beck et al., 2016) found that the ORCHIDEE model performed best in cold regions of the globe, which are not present in Australia, whereas it underestimated runoff for the other climate zones. For the temperate climate zones, W3RA performed best for runoff simulations, while the AWRA-L model performed worst for runoff for the temperate climate regions.

For actual evapotranspiration, best performances in the tropical climate zones were obtained by both HTESSEL and ORCHIDEE, whereas the ensemble median performed worst in these climatic regimes. for the arid regions, PCR-GLOBWB obtained the highest scores after evaluating actual evapotranspiration simulations, while ORCHIDEE performed worst for the arid regions. The temperate climate zones are best represented by the HTESSEL, and the worst performances for these climatic regimes are obtained by AWRA-L. Schellekens et al. (2017) demonstrated that ORCHIDEE tend to overestimate actual evapotranspiration in high-ET climate zones and these climate zones mainly occupy the Australian continent. Based on earlier research (Beck et al., 2016 ; Schellekens et al., 2017) and this research, ORCHIDEE seems to have difficulties in modeling water limited environments.

For soil moisture, W3RA performed best in the tropical climate zones, while HTESSEL simulations resulted in the worst performances for the tropical areas. For both the arid and the temperate climate zones, the ensemble median obtained the best results. AWRA-L performed worst for the arid climate zones and W3RA performed worst for the temperate climate zones. This research demonstrated that performances for GHMs (PCR-GLOBWB and W3RA) are better in arid climate zones than LSMs after evaluating actual evapotranspiration fields. However, for both runoff and soil moisture fields, differences in performances between the large-scale hydrological models were clearly visible, but not related to whether a model was a GHM or a LSM.

The difference in scales between gridded estimates for the large-scale hydrological models and the point observations could be a potential source for discrepancies between model simulations and observations. This holds especially for the comparison of soil moisture and actual evapotranspiration as streamflow observations are the integrated response of hydrological processes over a certain area. For soil moisture, which has a high spatiotemporal variability and which responds non-linear to associated physics (Brocca et al., 2010), this issue has been partly overcome by removing the bias. As a result, comparisons are based on soil moisture dynamics rather than absolute values (Orth et al., 2015) and earlier research found that these soil moisture dynamics are representative for a larger area around a certain measurement point (Mittelbach & Seneviratne, 2012). For comparison with the large-scale hydrological models, the AWRA-L has been upscaled. According to (Sablok & Aziz, 2008), upscaling leads to loss of information. In fact, the variability ratio between simulations and observations, which is evaluated by the KGE, is primarily affected by averaging of modeled output fields (Wainwright & Mulligan, 2012). In this research, low KGE (-bias) values have been found after evaluation of soil moisture fields by AWRA-L. However, high correlation values are obtained by AWRA-L for the same locations, which underpins this statement. Important to note, AWRA-L has been forced with a different dataset than the large-scale hydrological models. AWRA-L is forced with BAWAP at 0.05° resolution, while the large scale hydrological models are forced by WFDEI at 0.5° resolution.

In order to reduce parameter uncertainty and increase reliability of runoff fields, PCR-GLOBWB has been calibrated using four catchments in Australia. This research used a global calibration strategy, in which one parameter set is sought for multiple climate zones. K_{sat}, J and $StorCap$ have been chosen as effective parameters and the calibrated parameter set is $f_k = -0.25$, $f_j = 1.0$ and $f_s = 1.0$. To test whether PCR-GLOBWB is succesfully calibrated, validation has been carried out for all other catchments in Australia.

The calibration of PCR-GLOBWB has led to an increase in model performance for all climate zones. First of all, in the tropical climate zone, only 55% instead of 90% of the simulations performed unsatisfactory, which is equally good as the ensemble median. Furthermore, 15-20% of the simulations performed good in this climate zone instead of none for the default parameter setting and this percentage is even higher than the ensemble median. The improvements for the arid climate zone was less pronounced, but the PCR-GLOBWB model obtained 10% less unsatisfying simulations compared to the uncalibrated run. For the temperate climate zone, the amount of unsatisfying performances reduced with 10% by calibrating PCR-GLOBWB. Also, an increase of 10% for good performances compared with default PCRGLOB-WB run in this climate zone are achieved by calibrating PCR-GLOBWB model.

In this research, during calibration of PCR-GLOBWB, a bias in PCR-GLOBWB model performance was found for $f_k = -0.25$ and $f_s = 0.75$. In fact, the parameter prefactor $f_s = 0.75$ has been identified, but $f_s = 1.0$ is used for calibration as we investigated only the highest average KGE among the four selected catchments. Therefore, it is important to note that the parameter prefactor combination $f_j = -0.5$, $f_k = -0.25$, $f_s = 0.75$ could have led to increased reliability of PCR-GLOBWB runoff fields due to the equifinality principle, where multiple parameter combinations could lead to satisfying results. Also, minor improvements have been made for the arid climate zone. During calibration of PCR-GLOBWB, all model runs for the catchment located in the arid climate showed equally bad performance. So, for this climate zone, the bad performances are either related to a wrong selection of effective parameters or related to a poor model structure or poor forcing data rather than having a sub-optimal parameter combination for the selected effective parameters. According to Weedon et al. (2014), the combination of sub-grid variability neglection in the forcing of the models and rain gauges located in valleys only leads to substantial underestimation of rainfall in mountainous regions. So, improvements in forcing data at 0.5° but preferably at finer resolution could be valuable in order to enhance reliability of large-scale hydrological models (Weiland et al. 2015). Also, improvements in large-scale hydrological model performance can be achieved by using different forcing datasets (Lopez et al., 2017 ; Mizukami et al. 2013). Often, local calibration leads to better performances than global calibration (Gaborit, 2015), the global calibration method used in this research is promising as this technique imposes spatial consistency to the parameters, which enhances the model's applicability for multiple climate zones. Moreover, the global calibration produces a better temporal robustness (changing climate) than local calibration (Gaborit, 2015).

Furthermore, the selected calibrated parameter set in this study could be a local optimum as this research used the brute-force calibration technique, in which the parameter space is not visited completely during calibration. Possible improvements can be achieved by using the Shuffled Complex Evolution (SCE) global optimization algorithm, where a global optimum can reliably be found (Vrugt et al., 2003). Multiple studies have proven that the SCE global optimization algorithm is efficient, effective and consistent in finding the optimal model parameters for a hydrological model (Hogue et al., 2000 ; Boyle et al., 2000 ; Sorooshian et al., 1993). However, for complex hydrological models and for large areas, this algorithm is still computationally too intensive (Sharma et al., 2006). However, a parallel version of this algorithm managed to reduce the computation time required for automatic calibration of SWAT (Swayne et al., 2006), which is promising for further research.

Both for future water management purposes and to cope with natural hazards (droughts, floods) in the future, it is very important to gain insight in how models perform over different climate zones. This research aimed to fill this gap in the existing literature. This research demonstrated that large-scale hydrological models perform equally good er even better than the upscaled AWRA-L model, which served as benchmark. Also, this research showed that the calibration of PCR-GLOBWB has led to a substantial increase in the reliability of runoff fields. As a result, this research might have implications for developing countries, where a national modeling systems is absent. These countries rely on large-scale hydrological modeling and benefit from these evaluation studies. However, it is important to keep the purpose of modeling in mind before applying a certain large-scale hydrological model for a specific climate zone. In a flood prone region, reliable streamflow predictions are of paramount importance, whereas in a farming area reliable actual evapotranspiration simulations are more important for implementing potential irrigation measures. This is also related to the variable selection for calibration of a hydrological model for a specific area as this research demonstrated that the calibration of a large-scale hydrological model for a certain hydrological variable goes at the expense of model performance for other hydrological variables. This thesis focused on gauged catchments only. However, step-wise calibration of the PCR-GLOBWB model to for example satellite products as GLEAM actual evapotranspiration and ESA CCI soil moisture is very valuable for ungauged catchments as it opens the possibility to improve reliability of streamflow simulations without having streamflow observations to calibrate the model. Step-wise calibration is promising as it enhances the applicability of large-scale hydrological models for areas, where data records are absent. Step-wise calibration procedure has been done in earlier researches (Sutanudjaja et al., 2014 ; Lopez et al., 2017). So, based on this research, large-scale hydrological models, especially after calibration and validation, could be a valuable source for developing countries without a national hydrological model. Moreover, the global calibration approach used in this study is promising as it produces spatial consistency of parameters and temporal robustness, which enhances applicability and parameter stability of calibrated hydrological models.

5 Conclusion

This thesis aimed to evaluate Tier-1 EartH2Observe runoff, soil moisture and actual evapotranspiration simulations for four uncalibrated large-scale hydrological models in Australia. Based on this research, several conclusions can be drawn:

- For actual evapotranspiration fields, GHMs (especially PCR-GLOBWB) performed better in arid climate zones, whereas the LSMs obtrained higher scores in the tropical areas. For runoff and soil moisture simulations, differences in model performance was clearly visible, but this was not related to whether the model was a GHM or a LSM.

- The ensemble median performed as good or slighlty worse than the best large-scale hydrological after evaluating runoff, actual evapotranspiration and soil moisture simulations.

- The evaluated large-scale hydrological models performed similar or even better than the calibrated benchmark model, AWRA-L, for each evaluated hydrological variable. However, the upscaling procedure has led to detoriation of AWRA-L simulations.

- Validation of the calibrated PCR-GLOBWB model has led to model improvements for all climate zones compared to the reference scenario. Moreover, the calibrated PCR-GLOBWB values were for some climate zones higher than the multi-model ensemble median for that particular climate zone.

- this research proved that global scale hydrological model could be a valuable source of knowledge for developing countries without a fine resolution hydrological model. However, it is important to consider the purpose of modeling before calibrating and applying a large-scale hydrological model for a certain area as calibration of the large-scale hydrological model decreases model performance for other hydrological variables.

- performances for certain climate zones changed from very bad to bad by calibrating PCR-GLOBWB, which implies that insufficient simulations for those climate zones are due to poor model structure or poor forcing datasets instead of wrong model parameterization.

- Therefore, improvements in the forcing data globally at 0.5° or preferably at smaller scales is valuable in order to enhance reliability of large-scale hydrological models.

- Also, more research needs to be carried out for step-wise calibration in order to enhance the applicability of large-scale hydrological models. This is especially valuable for developing countries, where observational data is often lacking.

- The effect of other calibration techniques for PCR-GLOBWB should be investigated (e.g. the parallel SCE-UA global optimization algorithm: Swayne et al., 2006) The SCE-UA algorithm ensures to find a global optimum reliably and this parallel version reduced the computation time needed for the automatic calibration of complex models.

References

[1] N. K. Ajami, Q. Duan, S. Sorooshian, *An integrated hydrologic Bayesian multimodel combination framework: Confronting input, parameter and model structural uncertainty in hydrological prediction*, 2007, Water Resources Research, Vol. 43 (1).

[2] N. W. Arnell, *Thresholds and responses to climate change forcing: the water sector*, 2000, Climatic Change, Vol. 46 (305).

[3] N. W. Arnell, *Climate change and global water resources: SRES emissions and socio-economic scenarios,* 2004, Global Environmental Change, Vol. 14, 31-52.

[4] G. Balsamo, P. Viterbo, A. Beljaars, B. Van den Hurk, A. K. Betts, K. Scipal, *A revised hydrology for the ECMWF model: Verification from field site to terrestrial water storage and impact in the Integrated Forecast System,* 2009, Journal of Hydrometeorology, Vol. 10, 623-643.

[5] M. A. Bari, K. R. J. Smettem, M. Sivapalan, *Understanding changes in annual runoff following land use changes: a systematic data-based approach*, 2005, Hydrological Processes, Vol. 19, 2463-2479.

[6] H. E. Beck, A. I. J. M. van Dijk, A. de Roo, *Global maps of streamflow characteristics based on observations from several thousand catchments*, 2015, Journal of Hydrometeorology, Vol. 16, 1478-1501.

[7] H. E. Beck , A. I. J. M. van Dijk, A. de Roo, E. Dutra, G. Fink, R. Orth, J. Schellekens, *Global evaluation of runoff from ten state-of-the-art hydrological models*, 2016, Hydrology and Earth System Science, Vol. 21, 2881-2903.

[8] M. J. Best, M. Pryor, D. B. Clark, G. G. Rooney, R. L. H. Essery, C. B. Mnard, J. M. Edwards, M. A. Hendry, A. Porson, N. Gedney, L. M. Mercado, S. Sitch, E. Blyth, O. Boucher, P. M. Cox, C. S. B. Grimmond, R. J. Harding, *The Joint UK Land Environment Simulator (JULES), model description Part 1: Energy and water fluxes*, 2011, Geoscientific Model Development, Vol. 4, 677-699.

[9] K. Beven, A. Binley, *The future of distributed models: model calibration and uncertainty prediction*, 1992, Hydrological Processes, Vol. 6 (3), 279-298.

[10] K. J. Beven, *Prophecy, reality, and uncertainty in distributed hydrological modeling*, 1993, Advances in Water Resources, Vol. 16, 41-51.

[11] M. F. P., Bierkens, *Global hydrology 2015: state, trends, and directions*, 2015, Water Resources Research, Vol. 51, 4923-4947.

[12] BOM - Bureau of Meteorology, *Annual Climate Report*, 2015.

[13] D. P. Boyle, *Multicriteria calibration of hydrological models*, 2000, Ph.D. dissertation, Department of Hydrology and Water Resources, University of Arizona, Tucson.

[14] L. Brocca, F. Melone, T. Moramarco, R. Morbidelli, *Spatial-temporal variability of soil moisture and its estimation across scales*, 2010, Water Resources Research, Vol. 46 (2).

[15] L. Brocca, S. Hasenauer, T.Lacava, F. Melone, T. Moramarco, W. Wagner, W. Dorigo, P. Matgen, J. Martnez-Fernndez, P. Llorens, J. Latron, C. Martin, M. Bittelli, *Soil moisture estimation through ASCAT and AMSR-e sensors: an intercomparison and validation study across Europe*, 2011, Remote Sensing of Environment, Vol. 116, 3390-3408.

[16] D. B. Clark, L. M. Mercado, S. Sitch, C. D. Jones, N. Gedney, M. J. Best, M. Pryor, G. G. Rooney, R. L. H. Essery, E. Blyth, O. Boucher, R. J. Harding, C. Huntingford, P. M. Cox, *The Joint UK Land Environment Simulator (JULES), model description Part 2: Carbon fluxes and vegetation dynamics*, 2011, Geoscientific Model Development, Vol. 4, 701-722.

[17] W. T. Crow, E. F. Wood, *Impact of Soil Moisture Aggregation on Surface Energy Flux Prediction During SGP97*, 2002, Geophysical Research Letters, Vol. 29 (1).

[18] B. Decharme, R. Alkama, H. Douville, M. Becker, A. Cazenave, *Global evaluation of the ISBA-TRIP continental hydrological system. Part II:*

Uncertainties in river routing simulation related to flow velocity and groundwater storage, 2010, Journal of Hydrometeorology, Vol. 11, 601-617.

[19] B. Decharme, E. Martin, S. Faroux, *Reconciling soil thermal and hydrological lower boundary conditions in land surface models*, 2013, Journal of Geophysiscal Research: Atmosphere, Vol. 118, 7819-7834.

[20] D. Desilets, M. Zreda, T. P. A. Ferre, *Natures neutron probe: land surface hydrology at an elusive scale with cosmic rays*, 2010, Water Resources Research, Vol. 46 (11).

[21] W. P. A., van Deursen, *Geographical information systems and dynamic models: development and application of a prototype modelling language*, 1995.

[22] P.A. Dirmeyer, X. Gao, M. Zhao, Z. Guo, T. Oki, N. Hanasaki, *GSWP-2: Multimodel analysis and implications for our perception of the land surface*, 2006, Bulletin of the American Meteorological Society, Vol. 87, 1381-1397.

[23] P. Döll, K. Fiedler, J. Zhang, *Global-scale analysis of river flow alterations due to water withdrawals and reservoirs*, 2009, Hydrology and Earth System Science, Vol. 13, 2413-2432.

[24] Q. Duan et al., *Model Parameter Estimation Experiment (MO-PEX): An overview of science strategy and major results from the second and third workshops*, Journal of Hydrology, Vol. 320, 3-17.

[25] EartH2Observe, *Global Earth Observation for integrated water resource assessment: report on the current state-of-the-art water resources reanalysis*, 2015.

[26] M. Flörke, E. Kynast, I. Brlund, S. Eisner, F. Wimmer, J. Alcamo, *Domestic and industrial water uses of the past 60 years as a mirror of socio-economic development: A global simulation study*, 2013, Global Environmental Change, Vol. 23, 144-156.

[27] A. J. Frost, A. Ramchurn, M. Hafeez, F. Zhao, V. Haverd, J. Beringer,

P. Briggs, *Evaluation of AWRA-L: the Australian water resource Assessment model*, 2015, 21th International Congress on Modelling and Simulation, Gold Coast, Australia.

[28] E. Gaborit, S. Richard, S. Lachance-Cloutier, F. Anctil, R. Turcotte, *Comparing global and local calibration schemes from a differential split-sample test perspective*, 2015, Canadian Journal of Earth Sciences, 52 (11), 990-999.

[29] L. Gudmundsson et al., *Comparing Large-Scale Hydrological Model Simulations to Observed Runoff Percentiles in Europe*, 2012, Journal of Hydrometeorology, Vol. 13 (2), 604-620.

[30] A. Güntner, A. Bronstert, *Large-scale hydrological modelling of a semi-arid environment: model development, validation and application*, 2003, Global Change and Regional Impacts: Water Availability and Vulnerability of Ecosystems and Society in the Semiarid Northeast of Brazil, 217-228.

[31] I. Haddeland et al., *Multimodel estimate of the global terrestrial water balance: setup and first results*, 2011, Journal of Hydrometeorology, Vol. 12 (5), 869-884.

[32] M. Hafeez, A. Smith, A. Frost, S. Srikanthan, A. Elmahdi, J. Vaze, I. Prosser, *The bureaus operational AWRA modelling system in the context of Australian landscape and hydrological model products*, 2015, 36th Hydrology and Water Resources Symposium: The art and science of water, Barton, ACT: Engineers Australia, 1035-1042.

[33] S. Hagemann, L. Dumenil, *A Parameterization of the lateral waterflow for the global scale*, 1997, Climate Dynamics, Vol. 14, 17-31.

[34] W. R. Hamon: *Estimating Potential Evapotranspiration*, 1961, Journal of Hydraulics Division, Proceedings of the American Society of Civil Engineers, Vol. 87, 107-120.

[35] B. Hamraz, A. Akbarpour, M. P. Bilondi, S. S. Tableas, *On the assessment of ground water parameter uncertainty over an arid aquifer*, 2015, Arabian Journal of Geosciences, Vol. 8 (12), 10759-10773.

[36] H. J. E. Rodda, M. A. Little, *Understanding Mathematical and Statistical Techniques In Hydrology*, 2015.

[37] A. Hawdon, D. McJannet, J. Wallace, *Calibration and correction procedures for cosmic-ray neutron soil moisture probes located across Australia*, 2014, Water Resources Research, Vol. 50, 5029-5043.

[38] X. He, A. L. Hojberg, F. Jorgensen, J. C. Refsgaard, *Assessing hydrological model predictive uncertainty using stochastically generated geological models: Hydrological modelling using Stochastic Geological Models*, 2015, Hydrological Processes, Vol. 29 (19), 4293-4311.

[39] T. S. Hogue, S. Sorooshian, H. V. Gupta, A. Holz, and D. Braatz, *A multistep automatic calibration scheme for river forecasting models*, 2000, Journal of Hydrometeorology, Vol. 1, 524542.

[40] C. M. Holgate et al., *Comparison of remotely sensed and modelled soil moisture data sets across Australia*, 2016, Remote sensing of environment, Vol. 186, 479-500.

[41] T. J. Jackson et al., *Validation of soil moisture and ocean salinity (SMOS) soil moisture over watershed networks in the U.S.*, 2012, IEEE Transactions on Geoscience and Remote Sensing, Vol. 50 (5), 1530-1543

[42] A. Kauffeldt, F. Wetterhall, F. Pappenberger, P. Salamon, J. Thielen, *Technical review of large-scale hydrological models for the implementation in operational flood forecasting schemes on continental level*, 2016, environmental modeling and software, Vol. 75, 68-75.

[43] P. Krause, D. P. Boyle, F. Base, *Comparison of different efficiency criteria for hydrological model assessment*, 2005, Advances in Geosciences, Vol. 5, 89-97.

[44] G. Krinner, N. N. Viovy, N. de Noblet-Ducoudr, J. Oge, J. Polcher, P. Friedlingstein, P. Ciais, S. Stich, I. C. Prentice, *A dynamic global vegetation model for studies of the coupled atmosphere-biosphere system*, Global Biogeochemical Cycles, Vol. 19 (1).

[45] D. Lagrava, O. Malaspinas, J. Latt, B. Chopard, *Advances in multi-*

domain lattice Boltzmann grid refinement, 2012, Journal of Compututational Physics, Vol. 231, 4808-4822.

[46] C. Li, H. Wang, J. Liu, D. Yan, F. Lu, L. Zhang, *Effect of calibration data series length on performance and optimal parameters of hydrological model*, 2010, Water Science and Engineering, Vol. 3(4), 378-393.

[47] G. Lindstrom, B. Johansson, M. Persson, M. Gardelin, S. Bergstrom, *Development and test of the distributed HBV-96 hydrological model*, 1997, Journal of Hydrology, Vol. 201, 272-288.

[48] X. Liang, E. F. Wood, D. P. Lettenmaier, *Modeling ground heat flux in land surface parameterization schemes*, 1999, Journal of Geophysical Research, Vol. 104 (D8), 9581-9600.

[49] X. Liu et al., *Multimodel uncertainty changes in simulated river flows induced by human impact parameterizations*, 2007, Environmental Research Letters, Vol. 12 (2).

[50] D. Lohmann et al., *Streamflow and water balance intercomparisons of four land surface models in the North American Land Data Assimilation System project*, 2004, Journal of Geophysical Research, Vol. 109 (D7).

[51] P. Lopez, N. Wanders, J. Schellekens, L. J. Renzullo, E. H. Sutanudjaja, M. F. P. Bierkens, *Improved large-scale hydrological modelling through the assimilation of streamflow and downscaled satellite soil moisture observations*, 2016, Hydrology and Earth System Sciences, Vol. 20, 3059-3076.

[52] P. L. Lopez, E. Sutanudjaja, J. Schellekens, G. Sterk, M. Bierkens, *Calibration of a large-scale hydrological model using satellite-based soil moisture and evapotranspiration products*, 2017, Hydrology and Earth System Science, Vol. 21, 3125-3144.

[53] D. Lu, M. Ye, M. C. Hill, E. P. Poeter, G. P. Curtis, *A computer program for uncertainty analysis integrating regression and Bayesian methods*, 2014, Environmental Modelling & Software, Vol. 60, 45-56.

[54] J. R. Miller, G. L. Russel, G. Caliri, *Continental-scale river flow in climate models*, 1994, Journal of Climate, Vol. 7, 914-928.

[55] H. Mittelbach, S. I. Seneviratne, *A new perspective on the spatio-temporal variability of soil moisture: temporal dynamics versus time-invariant contributions*, 2012, Hydrology and Earth System Sciences, Vol. 16, 2169-2179.

[56] N. Mizukami, A. G. Slater, L. D. Brekke, M. M. Elsner, J. R. Arnold, S. Gangopadhyay, *Hydrologic Implications of Different Large-Scale Meteorological Model Forcing Datasets in Mountainous Regions*, 2013, Journal of Hydrometeorology, Vol.15, 474-488.

[57] J. M. Murphy, D. M. H. Sexton, D. N. Barnett, G. S. Jones, M. J. Webb, M. Collins, D. A. Stainforth, *Quantification of modelling uncertainties in a large ensemble of climate change simulations*, 2004, Nature, Vol. 430, 768-772.

[58] I. Mutenyo, A. P. Nejadhashemi, S. A. Woznicki, S. Giri, *Evaluation of SWAT Performance on a Mountainous Watershed in Tropical Africa*, 2013, Hydrology: Current Research, S14:001.

[59] T. Ngo-Duc, K. Laval, G. Ramillien, J. Polcher, A. Cazenave, *Validation of the land water storage simulated by Organising Carbon and Hydrology in Dynamic Ecosystems (ORCHIDEE) with Gravity Recovery and Climate Experiment (GRACE) data*, 2007, Water Resources Research, Vol. 43 (4).

[60] B. Nijssen et al., *Simulation of high latitude hydrological processes in the Torne-Kalix basin: PILPS phase 2(e) 2: Comparison of model results with observations*, 2003, Global Planetary Change, Vol. 38, 31-53.

[61] T. Oki, Y. Agata, S. Kanae, T. Saruhashi, D. Yang, K. Musiake, *Global assessment of current water resources using total runoff integrating pathways*, 2001, Hydrological Sciences Journal, Vol. 46, 983-995.

[62] R. Orth, S. I. Seneviratne, *Predictability of soil moisture and streamflow on subseasonal timescales: A case study*, 2013, Journal of Geophysical Research: Atmospheres, Vol. 118, 10963-79.

[63] R. Orth, S. I. Seneviratne, *Introduction of a simple-model based land surface dataset*, 2015, Environmental Research Letters, Vol. 10 (4).

[64] S. O. Owuor, K. Butterbach-Bahl, A. C. Guzha, M. C. Rufino, D. E. Pelster, E. Daz-Pins, L. Breuer *Groundwater recharge rates and surface runoff response to land use and land cover changes in semi-arid environments*, 2016, Ecological Processes, Vol. 5 (1).

[65] A. E. Raftery, T. Gneiting, F. Balabdaoui, and M. Polakowski, *Using Bayesian model averaging to calibrate forecast ensembles*, 2005, Monthly Weather Review, Vol. 133, 1155-1174.

[66] C. A. Reynolds, T. J. Jackson, W. J. Rawls, *Estimating soil water-holding capacities by linking the Food and Agriculture Organization Soil map of the world with global pedon daTableases and continuous pedotransfer functions*, 2000, Water Resources Research, Vol. 36, 3653-3662.

[67] R. Sablok, K. Aziz, *Upscaling and Discretization Errors in Reservoir Simulation*, 2008, Petroleum Science and Technology, Vol. 26, 1161-1186.

[68] V. Sharma, D. A. Swayne, D. Lam, W. Schertzer, *Parallel Shuffled Complex Evolution Algorithm for Calibration of Hydrological Models*, 2006, Conference paper, Proceedings of the 20th International Symposium on High-Performance Computing in an Advanced Collaborative Environment (HPCS'06).

[69] R. Shrestha, Y. Tachikawa, K. Takara, emphInput data resolution analysis for distributed hydrological modeling, 2006, Journal of hydrology, Vol. 319, 36-50.

[70] A. Sood, V. Smakhtin, *Global hydrological models: a review*, 2015, Hydrological Sciences Journal, Vol. 60, 549-565.

[71] S. Sorooshian, Q. Duan, V. K. Gupta, *Calibration of rainfall-runoff models: Application of global optimization to the Sacramento Soil Moisture accounting model*, 1993, Water Resources Research, Vol. 29, 11851194.

[72] C. A. Spruill, S. R. Workman, J. L. Taraba, *Simulation of daily and monthly stream discharge from small watersheds using the SWAT model*, 2000, American Society of Agricultural and Biological Engineers, Vol. 43, 1431-1439.

[73] E. Sutanudjaja, L. P. van Beek, N. Drost, I. E. M de Graaf, K. de Jong, M. W. Straatsma, Y. Wada, D. Wisser, M.F. Bierkens, *PCR-GLOBWB version 2.0: A High Resolution Integrated Global Hydrology and Water Resources Model*, 2014, American Geophysical Union, Fall Meeting.

[74] E. H. Sutanudjaja, L. P. H. van Beek, S. M. de Jong, F. C. van Geer and M. F. P. Bierkens, *Calibrating a large-extent high-resolution coupled groundwater-land surface model using soil moisture and discharge data*, 2014, Water Resources Research, Vol. 50, 687-705.

[75] E. Sutanudjaja, L. P. H. van Beek, Y. Wada, J. Bosmans, N. Drost, I. de Graaf, K. de Jong, P. Lopez Lopez, S. Pessenteiner, S. Oliver, M. Straatsma, N. Wanders, D. Wisser, M. Bierkens, *PCR-GLOBWB model*, 2016, Zenodo.

[76] P. Trambauer, S. Maskey, S. Winsemius, M. Werner, S. Uhlenbrook, *A review of continental scale hydrological models and their suitability for drought forecasting in (sub-saharan) Africa*, 2013, Physics and Chemistry of the Earth, Vol. 66, 16-26.

[77] K. Trenberth, *Conceptual framework for changes of extremes of the hydrological cycle with climate change*, 1999, Climate Change, Vol. 42, 327-339.

[78] L. P. H. van Beek, M. F. P. Bierkens, *The Global Hydrological Model PCR-GLOBWB: Conceptualization, Parameterization and Verification*, 2009, Technical Report, Department of Physical Geography, Utrecht University, Utrecht, The Netherlands.

[79] L. P. H. van Beek, Y. Wada, M. F. P. Bierkens, *Global monthly water stress: 1. Water balance and water availability*, 2011, Water Resources Research, Vol. 47 (7).

[80] J. M. Van Der Knijff, J. Younis, A. P. J. De Roo, *LISFLOOD: a GIS-based distributed model for river basin scale water balance and flood simulation*, 2010, International Journal Geographical Information Science, Vol. 24, 189-212.

[81] A. I. J. M. van Dijk, *Climate and terrain factors explaining streamflow*

response and recession in Australian catchments, 2010a, Hydrology and Earth System Sciences, Vol. 14, 159-169.

[82] A. I. J. M. van Dijk, *AWRA technical report 3, landscape model (version 0.5) technical description, WIRADA/CSIRO water for a healthy country flagship*, 2010b, Canberra.

[83] A. I. J. M. van Dijk, *Selection of an approximately simple storm runoff model*, 2010c, Hydrology and Earth System Sciences, Vol. 14, 447-458.

[84] A. I. J. M. van Dijk, L. A. Bruijnzeel, *modelling rainfall interception by vegetation of variable density using an adapted analytical model. Part 1 model description*, 2001, Journal of Hydrology, Vol. 247, 230-238.

[85] N. Viney, J. Vaze, R. Crosbie, B. Wang, W. Dawes, A. Frost, *AWRA-L v5.0: Technical description of model algorithms and inputs*, 2015, CSIRO, Australia.

[86] J. A. Vrugt, H. V. Gupta, W. Bouten, S. Sorooshian, *A Shuffled Complex Evolution Metropolis algorithm for optimization and uncertainty assessment of hydrologic model parameters*, 2003, Water Resources Research, Vol. 39 (8).

[87] Y. Wada, D. Wisser, M. F. P. Bierkens, *Global modeling of withdrawal, allocation and consumptive use of surface water and groundwater resources*, 2014, Earth System Dynamics, Vol. 5, 15-40.

[88] J. Wainwrigth, M. Mulligan, *Environmental modelling: Finding Simplicity in Complexity*, 2012.

[89] K. Wang, Z. Li, M. Cribb, *Estimation of evaporative fraction from a combination of day and night land surface temperatures and NDVI: A new method to determine the Priestley-Taylor parameter*, 2006, Remote Sensing of Environment, Vol. 102, 293-305.

[90] W. E. Walker, P. Harremos, J. Rotmans, J.P. van der Sluijs, M.B.A. van Asselt, P. Janssen, M. P. Krayer van Krauss, *Defining uncertainty. a conceptual basis for uncertainty management in model-based decision support*, 2003, Integrated Assessment, Vol. 4, 5-17.

[91] G. P. Weedon, G. Balsamo, N. Bellouin, S. Gomes, M. J. Best, P. Viterbo, *The WFDEI meteorological forcing data set: WATCH Forcing Data methodology applied to ERA-Interim reanalysis data*, 2014, Water Resources Research, Vol. 50 (9), 7505-7514.

[92] F. C. Sperna Weiland, J. A. Vrugt, L. P. H. van Beek, A. H. Weerts, M. F. P. Bierkens, *Significant uncertainty in global scale hydrological modeling from precipitation data errors*, 2015, Journal of Hydrology, Vol. 529, 1095-1115.

[93] R. L. Wilby, S. Dessai, *Robust adaptation to climate change*, 2010, Vol. 65 (7), 180-185.

[94] World Meteorological Organization, *Comprehensive assessment of the freshwater resources of the world*, 1997, Stockholm, Sweden.

[95] E. F. Wood et al., *The project for intercomparison of land-surface parameterization schemes (PILPS) Phase-2(c) Red-Arkansas River basin experiment: 1. Experiment description and summary intercomparisons*, 1998, Global Planetary Change, 19 (1-4), 115-135.

[96] H. Wu, Z.L. Li, *Scale issues in remote sensing: a review on analysis, processing and modeling, Sensors*, 2009, Vol. 9, 1768-1793.

[97] D. Yamazaki, T. Oki, S. Kanae, *Deriving a global river network map and its sub-grid topographic charateristics from a fine-resolution flow direction map*, 2009, Hydrology and Earth System Sciences, Vol. 13, 2241-2251.

[98] M. Yebra, A. van Dijk, R. Leuning, A. Huete, J. P. Guerschman, *Evaluation of optical remote sensing to estimate actual evapotranspiration and canopy conductance*, 2013, Remote Sensing Environment, Vol. 129, 250-261.

[99] K. K. Yilmaz, H. V. Gupta, T. Wagener, *A process-based diagnostic approach to model evaluation: Application to the NWS distributed hydrologic model*, 2008, Water Resources Research, Vol. 44 (9).

[100] W. Zhao, A. Li, *A Review on Land Surface Processes Modelling over Complex Terrain*, 2015, Advances in Meteorology, Vol. 2015.

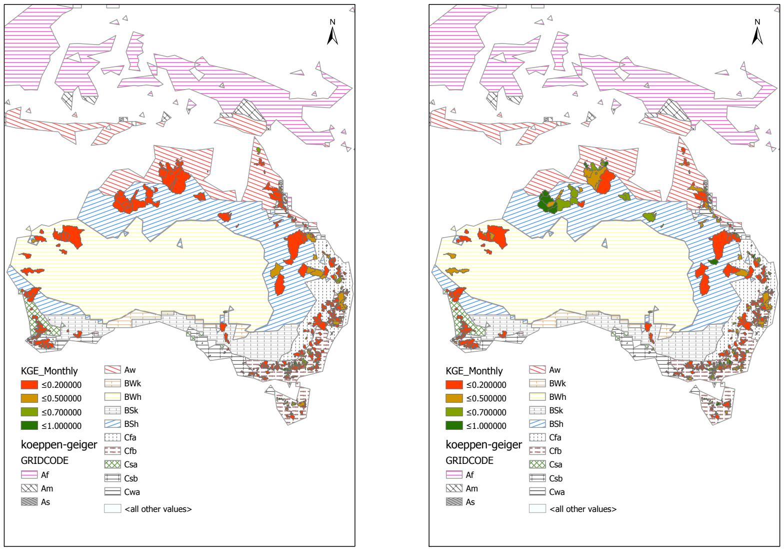
[101] Y. Zhang, N. Viney, A. Frost, A. Oke, M. Brooks, Y. Chen and N. Campbell, *Collation of Australian modellers streamflow dataset for 780 unregulated Australian catchments*, 2013, CSIRO: Water for a Healthy Country National Research Flagship.

[102] M. Zreda, D. Desilets, T. P. A. Ferre, R. L. Scott, *Measuring soil moisture content non-invasively at intermediate spatial scale using cosmic-ray neutrons*, 2008, Geophysical Research Letters, Vol. 35 (21).

A Figures and Tables

A.1 Runoff

A.1.1 Additional performance maps monthly simulations

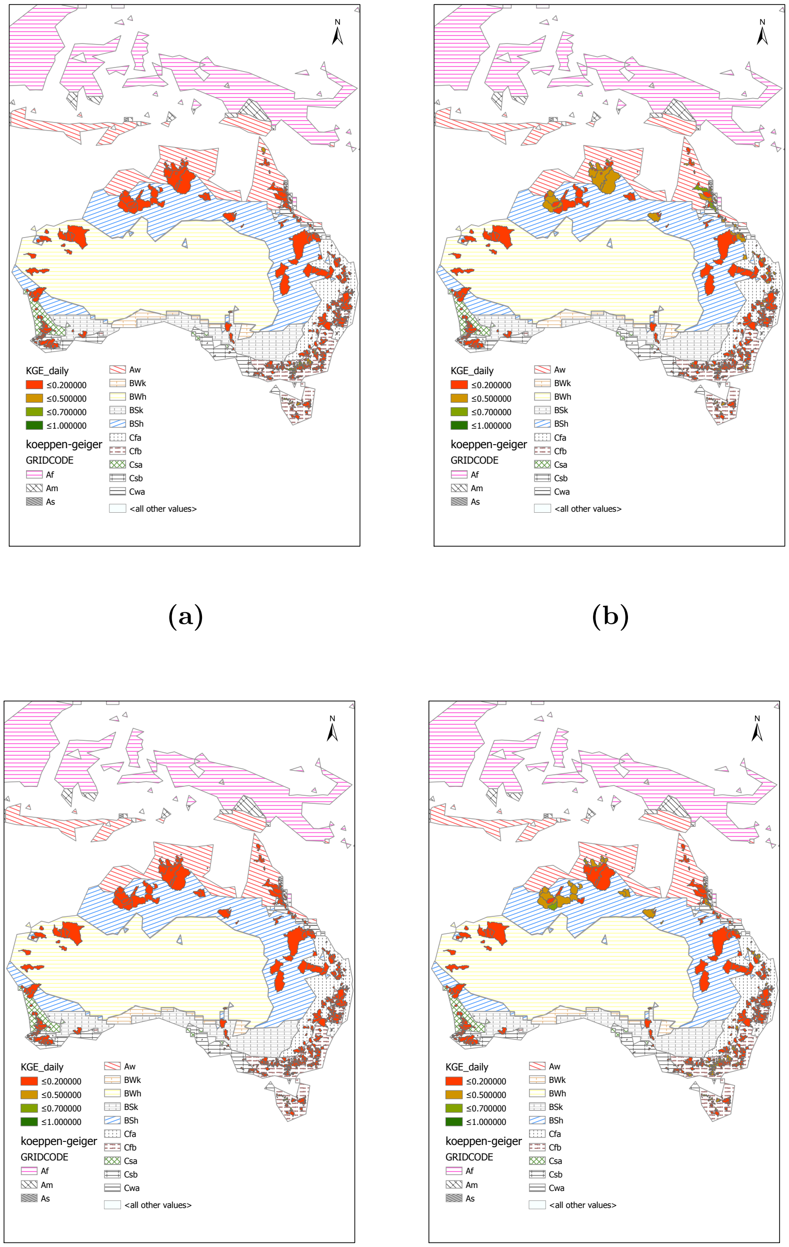


(a)

(b)

Fig. 15: Additional streamflow performance maps for PCR-GLOBWB (a) and W3RA (b). These performance maps are based on monthly simulations.

A.1.2 Performance maps daily simulations

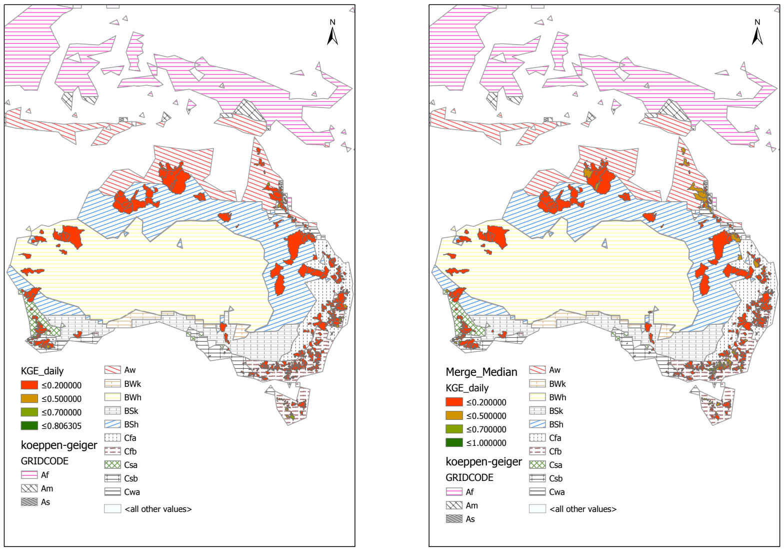


(a)

(b)

(c)

(d)

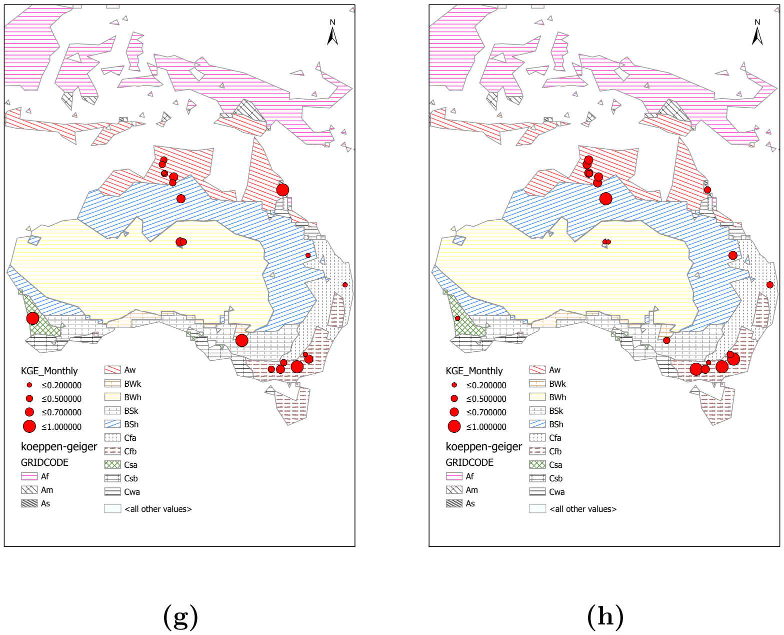


(e) **(f)**

Fig. 15: Streamflow performance maps for PCR-GLOBWB (a), HTESSEL (b), ORCHIDEE (c), W3RA (d), AWRA-L (e) and the ensemble median (f). These maps are based on daily simulations

A.2 Actual evapotranspiration

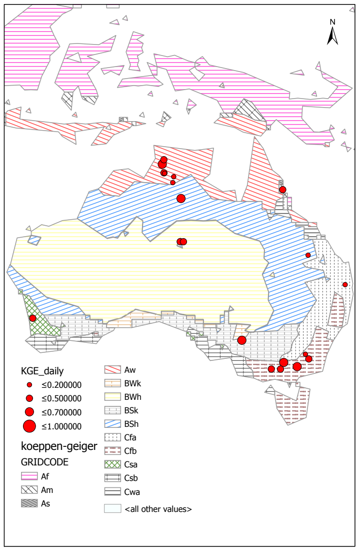
A.2.1 Additional performance maps monthly simulations



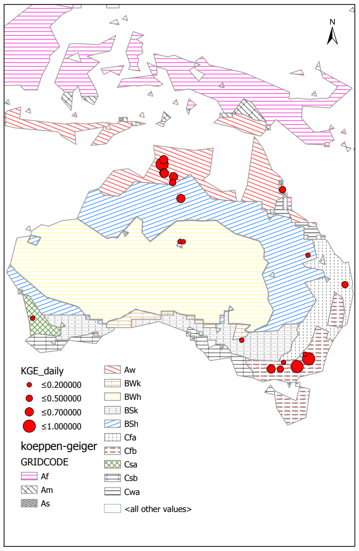
(g) (h)

Fig. 16: Performance maps for actual evapotranspiration for monthly estimates for PCR-GLOBWB(a) and W3RA (b).

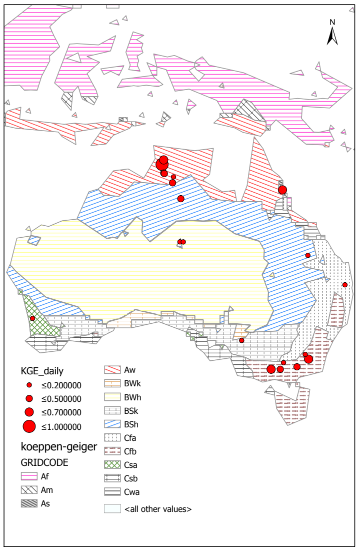
A.3 Performance maps daily simulations



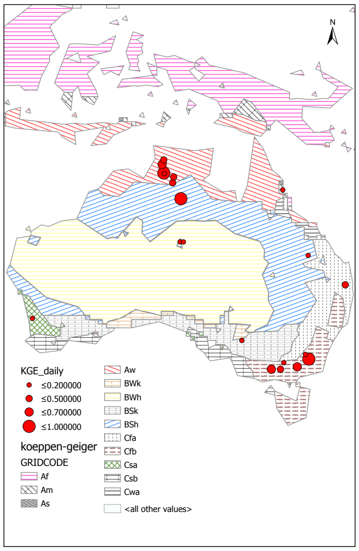
(a)



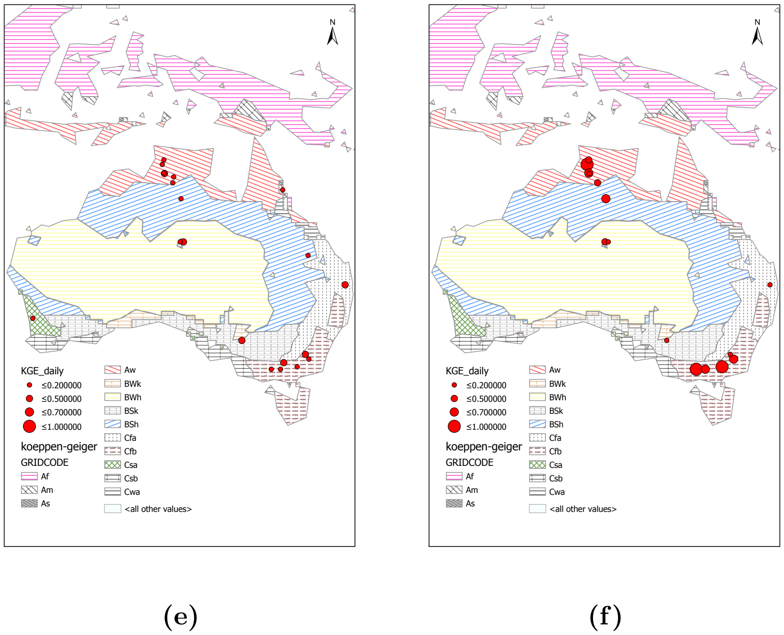
(b)



(c)



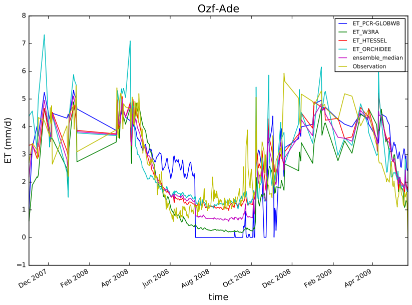
(d)



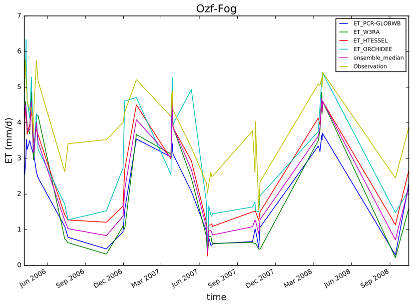
(e) (f)

Fig. 16: Performance maps for large-scale hydrological models (PCR-GLOBWB (a), HTESSEL (b), ORCHIDEE (c), W3RA (d), AWRA-L (e) and the ensemble median (f) for daily resolution actual evapotranspiration estimates.

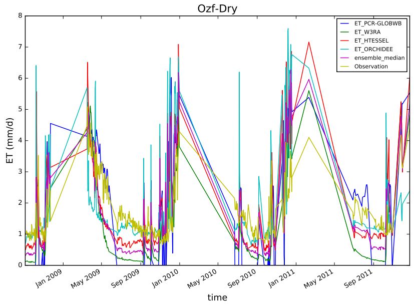
A.3.1 Actual evapotranspiration time series



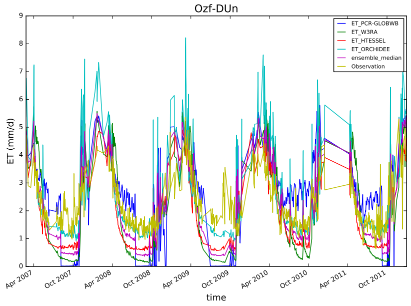
(g)



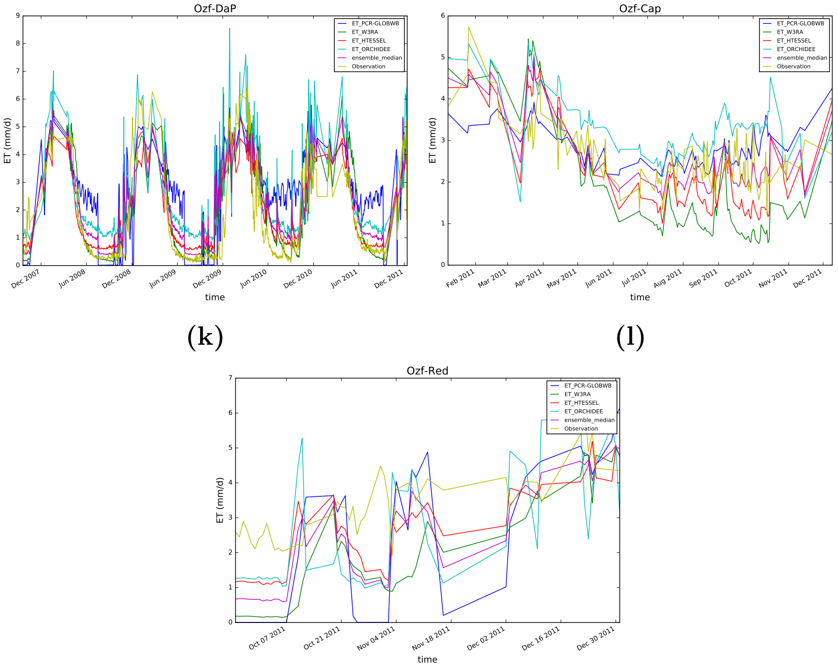
(h)



(i)



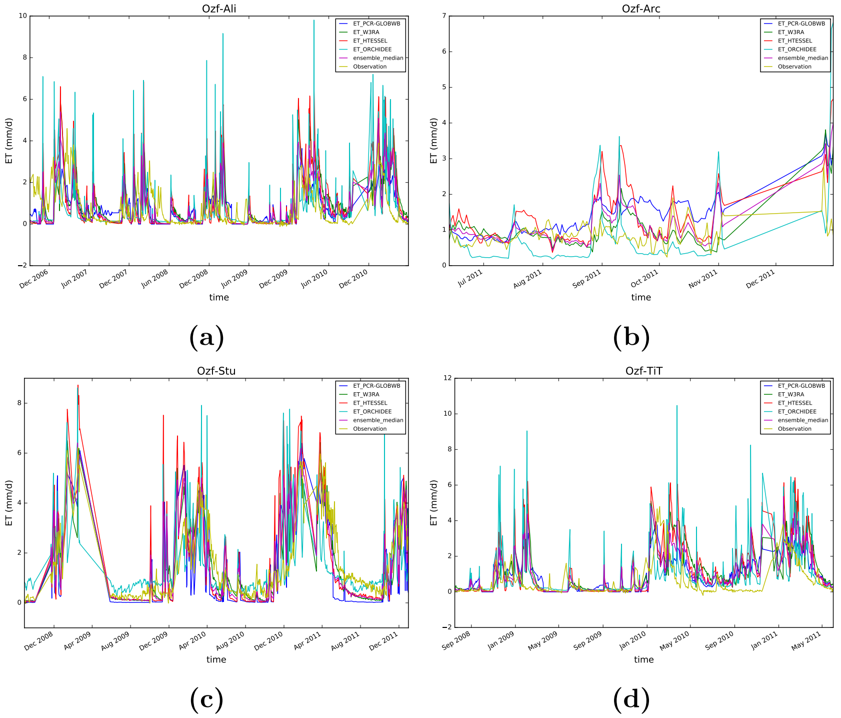
(j)



(k) **(l)**

(m)

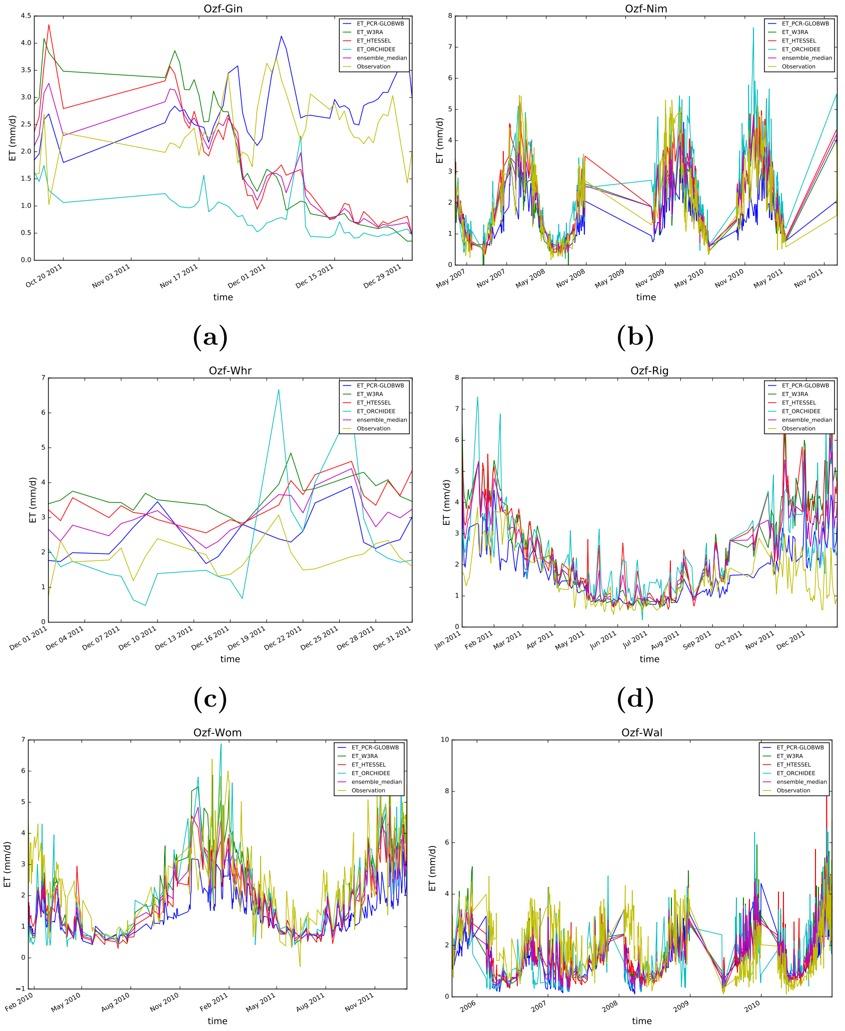
Fig. 16: Time series for modeled daily estimates of actual evapotranspiration, observed daily actual evapotranspiration, and the ensemble median of the large-scale hydrological models. These timeseries are generated for Ozflux sites located in the tropical climate-zones.



(a) **(b)**

(c) **(d)**

Fig. 17: Time series for modeled daily estimates of actual evapotranspiration, observed daily actual evapotranspiration, and the ensemble median of the large-scale hydrological models. These time series are generated for Ozflux sites located in the arid climate zones.


Ozf-Gin
ET (mm/d)
time
Oct 20 2011
Nov 03 2011
Nov 17 2011
Dec 01 2011
Dec 15 2011
Dec 29 2011
Ozf-Nim
May 2007
Nov 2007
May 2008
Nov 2008
May 2009
Nov 2009
May 2010
Nov 2010
May 2011
Nov 2011
Ozf-Whr
Dec 01 2011
Dec 04 2011
Dec 07 2011
Dec 10 2011
Dec 13 2011
Dec 16 2011
Dec 19 2011
Dec 22 2011
Dec 25 2011
Dec 28 2011
Dec 31 2011
Ozf-Rig
Jan 2011
Feb 2011
Mar 2011
Apr 2011
May 2011
Jun 2011
Jul 2011
Aug 2011
Sep 2011
Oct 2011
Nov 2011
Dec 2011
Ozf-Wom
Feb 2010
May 2010
Aug 2010
Nov 2010
Feb 2011
May 2011
Aug 2011
Nov 2011
Ozf-Wal
2006
2007
2008
2009
2010
ET_PCR-GLOBWB
ET_W3RA
ET_HTESSEL
ET_ORCHIDEE
ensemble_median
Observation

(a)

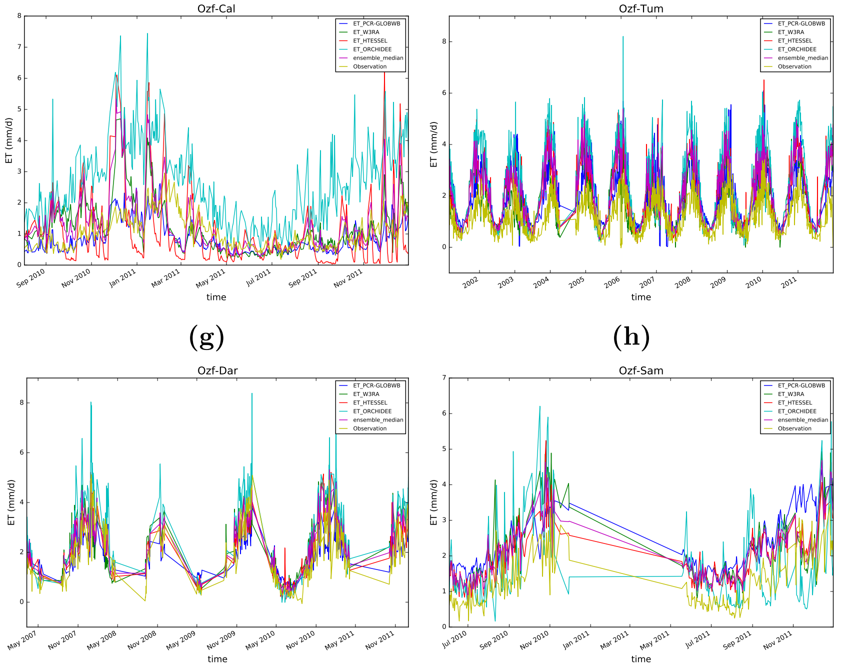
(b)

(c)

(d)

(e)

(f)



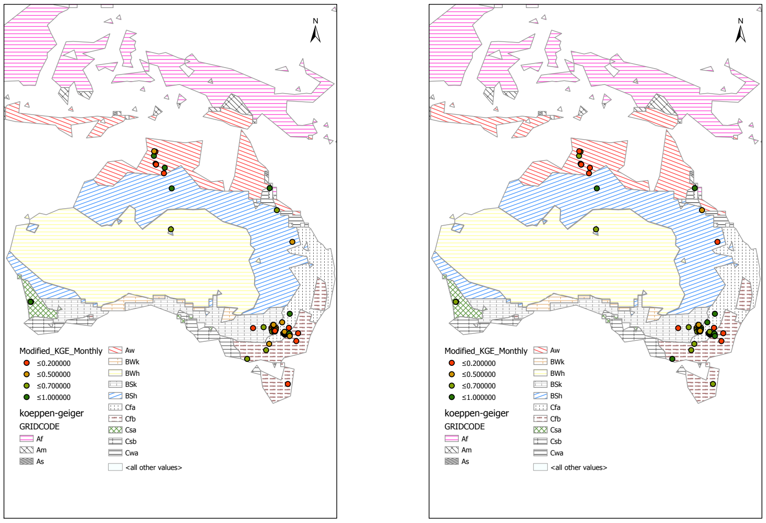
(g) **(h)**

(i) **(j)**

Fig. 17: Time series for modeled daily estimates of actual evapotranspiration, observed daily actual evapotranspiration, and the ensemble median of the large-scale hydrological models. These timeseries are generated for Ozflux sites located in the temperate climate-zones.

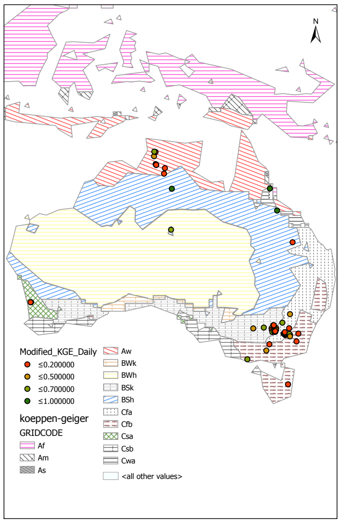
A.4 Soil moisture

A.4.1 Additional performance maps monthly simulations

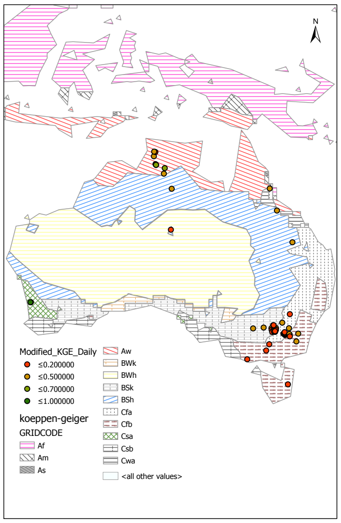


(k) (l)

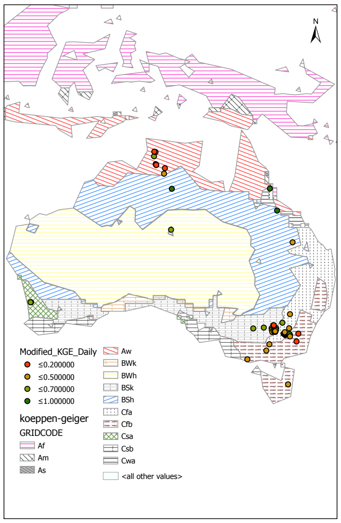
Fig. 18: Additional soil moisture performance maps for PCR-GLOBWB (a) and HTESSEL (b). These performance maps are based on monthly simulations.


N
Modified_KGE_Daily
≤0.200000
≤0.500000
≤0.700000
≤1.000000
koeppen-geiger
GRIDCODE
Af
Am
As
Aw
BWk
BWh
BSk
BSh
Cfa
Cfb
Csa
Csb
Cwa
<all other values>

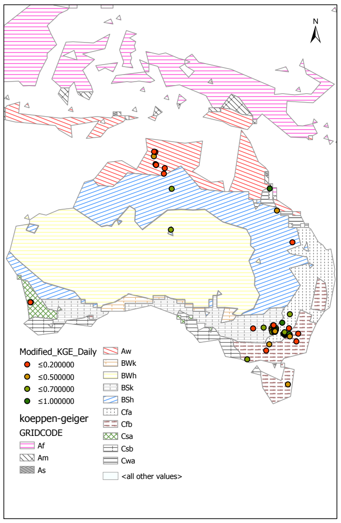
(a)


N
Modified_KGE_Daily
≤0.200000
≤0.500000
≤0.700000
≤1.000000
koeppen-geiger
GRIDCODE
Af
Am
As
Aw
BWk
BWh
BSk
BSh
Cfa
Cfb
Csa
Csb
Cwa
<all other values>

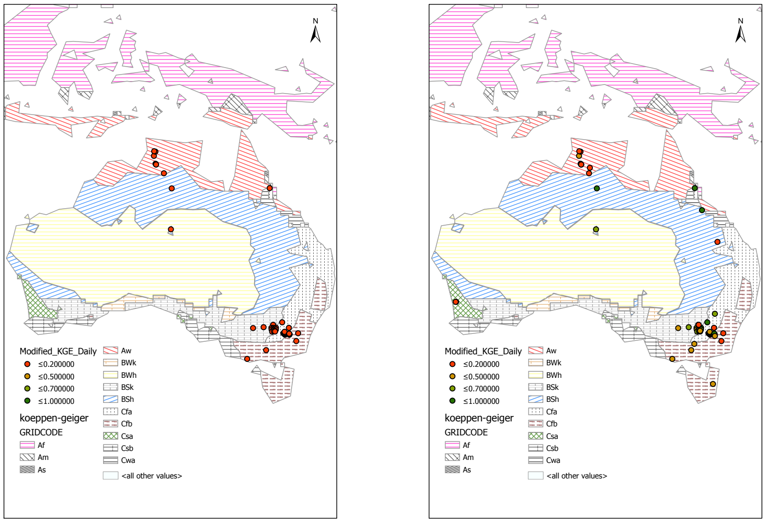
(b)


N
Modified_KGE_Daily
≤0.200000
≤0.500000
≤0.700000
≤1.000000
koeppen-geiger
GRIDCODE
Af
Am
As
Aw
BWk
BWh
BSk
BSh
Cfa
Cfb
Csa
Csb
Cwa
<all other values>

(c)


N
Modified_KGE_Daily
≤0.200000
≤0.500000
≤0.700000
≤1.000000
koeppen-geiger
GRIDCODE
Af
Am
As
Aw
BWk
BWh
BSk
BSh
Cfa
Cfb
Csa
Csb
Cwa
<all other values>

(d)



(e)

(f)

Fig. 18: Soil moisture performance maps for PCR-GLOBWB (a), W3RA (b), ORCHIDEE (c), HT-ESSEL (d), AWRA (e) and the ensemble median (e). These performance maps are based on daily simulations.

A.4.2 Performance maps daily simulations

B Suplementary Tables

Tab. 16: Pearsons correlation values for soil moisture simulations by the AWRA-L model. These values are based on monthly time resolution performance evaluation.

Site-Code	Climate-Zone	Pearsons monthly Correlation
M1	Cfb	0.60
M2	Cfb	0.69
M3	Cfa	0.80
M4	Cfa	0.83
M5	Bsk	0.87
M6	Bsk	0.85
M7	Bsk	0.79
Y1	Bsk	0.89
Y2	Bsk	0.92
Y3	Bsk	0.90
Y4	Bsk	0.90
Y5	Bsk	0.91
Y6	Bsk	0.64
Y7	Bsk	0.87
Y8	Bsk	0.92
Y9	Bsk	0.89
Y10	Bsk	0.86
Y11	Bsk	0.89
Y12	Bsk	0.90
Y13	Bsk	0.77
K1	Cfa	0.73
K2	Cfa	0.81
K3	Cfa	0.79
K4	Cfa	0.82
K5	Cfa	0.81
K6	Cfa	0.78
K7	Cfa	0.77
K8	Cfa	0.82
K10	Cfa	0.71
K11	Cfa	0.81
K12	Cfa	0.82
K13	Cfa	0.86
K14	Cfa	0.85
A1	Cfb	0.78
A2	Cfb	0.87
A3	Cfb	0.86
A4	Cfb	0.87
A5	Cfb	0.80
Site06	Bsh	0.99
Adelaide River	Cfb	0.87
Alice Springs Mulga	Cfb	0.93
Daly Pasture	Aw	0.94
Daly Uncleared	Aw	0.94
Dry River	Aw	0.94
Fogg Dam	Aw	0.09
Howard Springs	Aw	0.92
Sturt Plains	Bsh	0.88
Wallaby Creek	Cfb	0.44
Daly Regrowth	Aw	0.96
Otway	Cfb	0.88