Friedrich-Schiller University Jena
Institute of Geography
Department of Geoinformatics

Master Thesis

Macroscopic diagnostic modeling of the hydrological cycle: Understanding the dynamics of water pools in snow affected regions.

A thesis submitted in fulfillment of the requirements for the degree of Master of Science in Geoinformatics.

Author: Tina Trautmann
Matr. 120409
tina.trautmann@uni-jena.de

Supervisors: Dr. Manfred Fink
Friedrich-Schiller University Jena

Dr. Martin Jung
Max-Planck-Institute for
Biogeochemistry Jena

Submission date: 15.07.2016

Contents

List of Figures

List of Tables

List of Acronyms and Symbols

actET	scaled evapotranspiration [mm d^{-1}]
AIRS	Atmospheric Infrared Sounder (instrument onboard Aqua satellite)
AVHRR	Advanced Very High Resolution Radiometer (instrument onboard NOAA-7, MetOp-A satellites)
BATS	Biosphere-Atmosphere Transfer Scheme
BFGS	Broyden-Fletcher–Goldfarb-Shanno optimization method
BFI1-4	base flow index computed by using four different techniques
BRDF	bidirectional reflectance distribution function
C	carbon
Cal	grid cells used for final model calibration
CERES	Clouds and the Earth's Radiant Energy Systems (instrument onboard Terra and Aqua satellites)
CF1-6	Cost function 1 – 6
CMAES	Covariance Matrix Adaption Evolution Strategy
CO_2	carbondioxide
covar	covariance
c_p	specific heat of air at constant pressure [MJ kg^{-1} K^{-1}]
CRI	Coastline Resolution Improvement
CRI	Coastline Resolution Improvement
CRU	Climate Research Unit
dd	day-degree approach
DLR	German Aerospace Center
EB	energy balance
ECMWF	European Centre for Medium-Range Weather Forecasts
ED	Euclidian Distance
Eq.	equation

ERA-interim	ECMWF interim Reanalysis data
ESA	European Space Agency
ET	evapotranspiration [mm d^{-1}]
ET_{mod}	modeled evapotranspiration [mm d^{-1}]
ET_{obs}	observed evapotranspiration [mm d^{-1}]
ETSub	sublimation from snow cover [mm d^{-1}]
EU	European Union
Eval	grid cells used for model evaluation
FLUXNET	global network of micrometeorological tower sites
Fmincon	function minimization with constraints algorithm
fPAR	fraction of photosynthetic active radiation
FSC	fractional snow cover [-]
G	ground heat flux [MJ m^{-2} d^{-1}]
g_d	groundwater depletion factor [d^{-1}]
g_r	groundwater recharge rate [-]
GA	Genetic Algorithm
GCM	global circulation model
GEOS	Geosynchronous Operational Environmental Satellites, United States
GHGA	Guided Hybrid Genetic Algorithm
GHM	global hydrological model
GLEAM	Global Land-surface Evaporation: the Amsterdam Methodology
GPCC	Global Precipitation Climatology Centre
GPCP-1DD	Global Precipitation Climatology Project 1 Degree Daily precipitation product
GPCP-SG	Global Precipitation Climatology Project satellite-gauge combined monthly precipitation product
GPS	Global Positioning System
GRACE	Gravity Recovery and Climate Experiment
GSFmincon	global search algorithm applying Fmincon
GW	ground water storage [mm]

GWAVA	Global Water Availability Assessment model
HBV	Hydrologiska Byråns Vattenbalansavdelning model
HDTM 1.0	Hydro Dynamic Model
HM	hydrological model
HRU	hydrological response unit
H-TESSEL	revised land surface Hydrology of the Tiled ECMWF Scheme for Surface Exchanges over Land
HUT	Helsinki University of Technology
In	infiltration to soil /soil water recharge [mm d^{-1}]
ISBA-TRIP	Interactions between Soil, Biosphere, and Atmosphere – Total Runoff Integrating Pathways
IW	input water [mm d^{-1}]
JPL	Jet Propulsion Laboratory
K	base flow recession constant [d^{-1}]
KG	Köppen-Geiger climate zones
KGE	Kling-Gupta efficiency
KM12	simple water balance model by Koster and P. Mahanama (2012)
LaD	Land Dynamic model
LAI	leaf area index
lat	latitude
lb	lower parameter bounds
LC	land cover
LE	latent energy [MJ m^{-2} d^{-1}]
long	longitude
LPJ/ LPJmL	Lund-Potsdam-Jena/ managed Land model
LSM	land surface model
Lsqnonlin	least-square nonlinear search algorithm

M	melt from snow cover [mm d^{-1}]
m_r	radiation factor for snow melt [mm MJ^{-1}]
m_t	degree-day factor for snow melt [mm K^{-1}]
MA	ratio of molecular weight of water vapor/dry air
Macro-PDM	Macro Probability Model
mascon	mass concentration block
MATLAB	MATrix LABoratory software
MaxFunEval	maximum number of function evaluations
MaxIter	maximum number of iterations
MEF	Nash-Sutcliff efficiency
MERIS	Medium Resolution Imaging Spectormeter (instrument onboard Envisat)
Meteosat	Meteorological Satellite, European Community
MODIS	Moderate-resolution Imaging Spectroradiometer (instrument onboard Terra and Aqua satellites)
MPI-BGC	Max-Planck-Institute for Biogeochemistry, Jena
MPI-HM	MPI Hydrology Model
MSC	mean seasonal cycle
MSE	Mean Square Error
N	Number of data points
NaN	not a number
NASA	National Aeronautics and Space Administration, United States
NCEP	National Centers for Environmental Prediction
NH	Northern Hemisphere
P	daily precipitation [mm d^{-1}]
p_et	scaling parameter for evapotranspiration [-]
p_gd	scaling parameter for K in Beck-GW models
p_gr	scaling parameter for BFI in Beck-GW models
p_sf	scaling parameter for snowfall [-]
Pair	atmospheric pressure [kPa]

PBIAS	percentage bias
PCR-GLOBWB	PCRaster Global Water Balance model
Perc	percolation to groundwater /groundwater recharge [mm d^{-1}]
PFT	plant functional type
pSigma	step size of CMAES
Q	total discharge [mm d^{-1}]
q_t	recession time scale for Qs [d]
Qbase	base flow [mm d^{-1}]
Qdir	direct runoff [mm d^{-1}]
Qs	runoff from soil [mm d^{-1}]
R	range
r	Pearson's product-moment correlation coefficient
R^2	coefficient of determination
relHum	relative humidity [%]
RF	rainfall [mm d^{-1}]
rk	rank
RL↑	upward longwave radiation [MJ m^{-2} d^{-1}]
RL↓	downward longwave radiation [MJ m^{-2} d^{-1}]
RMSE	Root Mean Square Error
Rn	net radiation [MJ m^{-2} d^{-1}]
RS↑	upward shortwave radiation [MJ m^{-2} d^{-1}]
RS↓	downward shortwave radiation [MJ m^{-2} d^{-1}]
s_exp_berg	runoff coefficient [-]
s_exp_budyko	retention efficiency of soil [-]
s_exp_simple	form of relation between runoff and water storage [-]
s_fac_simple	tendency of soil to release water [-]
s_max	maximum soil water holding capacity [mm]
SA	Simulated Annealing algorithm

SAC-SMA	Sacramento Soil Moisture Accounting Model
SCE	Shuffled Complex Evolution algorithm
SeaWiFS	Sea-viewing Wide Field-of-View Sensor (instrument onboard Orb-View-2)
SF	snowfall [mm d^{-1}]
SHE	Systéme Hydrologique Européen model
SM	soil moisture / soil water storage [mm]
SMMR	Scanning Multichannel Microwave Radiometer (instrument onboard Nimbus-7)
sn_c	minimum SWE that ensures complete snow coverage of the grid cell [mm]
sn_α	sublimation resistance of snow [-]
SSMI	Special Sensor Microwave Imager (instrument onboard Defense Meteorological Satellite Program's satellites F8, F10, F11, F12, F13, F15)
SSMIS	Special Sensor Microwave Imager Sounder (instrument onboard Defense Meteorological Satellite Program's satellites F16, F17, F18)
SWAT	Soil and Water Assessment Tool model
SWE	snow water equivalent [mm d^{-1}]
SWE_{mod}	modeled snow water equivalent [mm]
SWE_{obs}	observed snow water equivalent [mm]
T	temperature [°C]
T_B	brightness temperature [K]
thrsh	threshold
TIN	triangulated irregular network
TMPI	Threshold-Match-Precipitation Index
TolFun	termination tolerances for the cost function
TolX	termination tolerances for the parameter scaling factors
TOPMODEL	Topographic Model
TOVS	Television and Infrared Observation Satellite Operational Vertical Sounder (High Resolution Infrared Radiation Sounder, Microwave

	Sounding Unit and Stratospheric Sounding Unit onboard TIROS satellites)
tRIB	TIN-based Real-Time Basin Simulator
TWS	total water storage [mm]
TWS_{mod}	modeled total water storage [mm]
TWS_{obs}	observed total water storage [mm]
ub	upper parameter bounds
varGW	variance of GW
varSM	variance of SM
varSWE	variance of SWE
varTWS	variance of TWS
varW	variance of W
VIC	Variable Infiltration Capacity model
W	total amount of available water [mm]
w mean rk	weighted mean rank
WAI	water availability index
WASMOD-M	Water and Snow balance Modeling system
WATCH	Water and Global Change Project of the EU
WaterGAP	Water – Global Analysis and Prognosis model
WB	water balance
WBM_{plus}	Water Balance Model plus Irrigation
WBM-WTM	Water Balance Model – Water Transport Model
WFDEI	WATCH forcing data methodology applied to the ERA-Interim reanalysis data
wMEF	weighted Nash-Sutcliff efficiency
wMSE	weighted Mean Square Error
x_{mod}	modeled variable
x_{obs}	observed variable

Greek Letters

γ	psychrometric constant [kPa K^{-1}]
Δ	slope of temperature/saturated vapor pressure curve [kPa K^{-1}]
ΔGW	change in GW
ΔTWS	change in TWS
λ	latent heat of sublimation of ice [MJ kg^{-1}]
μ_{mod}	mean of modeled variable
μ_{obs}	mean of observed variable
ρ	Spearman's rank correlation coefficient
σ	uncertainty
σ_{ET}	uncertainty of ET_{obs}
σ_{mod}	standard deviation of modeled variable
σ_{obs}	standard deviation of observed variable
σ_{SWE}	uncertainty of SWE_{obs}
σ_{TWS}	uncertainty of TWS_{obs}
α	variability error
β	bias error

Developed Model Variants

Beck1-4	Beck-GW models applying K and BFI1-BF4 estimates
Beck-BFI1-4	Beck-GW models applying BFI1-BF4 estimates
Beck-GW	groundwater model variants applying Beck et al. (2015) BFI and K estimates
Beck-K	Beck-GW model applying K estimates
BergBasic	model variant applying Bergström approach and basic runoff
BergGW	model variant applying Bergström approach and groundwater runoff
BudykoBasic	model variant applying Budyko approach and basic runoff
BudykoGW	model variant applying Budyko approach and groundwater runoff
SatBasic	model variant applying saturation approach and basic runoff
SatGW	model variant applying saturation approach and groundwater runoff
SimpleBasic	model variant applying simple approach and basic runoff

1 Introduction

In the continuing debate on climate change, the carbon cycle plays a major role as additional input of carbon gases into the atmosphere is seen among the main drivers for global warming (IPCC, 2014; Riebeek, 2011). Since the carbon reservoirs atmosphere, biosphere, soil and ocean are strongly interconnected, better comprehension on how these interactions affect and respond to global change is necessary (Harding et al., 2011; Hari et al., 2016; Tallaksen et al., 2015). In this context, research at the Max Planck Institute for Biogeochemistry, Jena (MPI-BGC) aims to develop and investigate methods and models to describe the dynamics of the terrestrial biosphere within the Earth system, focusing on states and fluxes related to the carbon cycle (MPI-BGC, 2016). However, the global carbon cycle is closely linked to the hydrological cycle (Battin et al., 2009; Döll et al., 2015; Falkowski et al., 2000), and thus integration of both into global simulation models is required (Gerten et al., 2004; Sood and Smakhtin, 2015). While the coupled terrestrial carbon and water cycle is simulated by dynamic global vegetation models (DGVMs), these models focus on climate-vegetation interactions and their hydrological performance often remains unverified (Gerten et al., 2004). On the other hand, hydrological models (HMs) usually serve for water management and flood forecasting (Gupta, 2011) and lack sufficient representation of vegetation and biosphere-hydrosphere interactions (Gerten et al., 2004). Additionally, the complexity of existing models, both DGVMs and HMs, increased continually over time (Döll et al., 2015) by implementing processes thought to be relevant, leading to the pitfall of possible model overparametrization (Jakeman and Hornberger, 1993). Thus, in general the ‘best’ model is still the one which gives results close to reality with the use of least parameters and model complexity (Sorooshian et al., 1993).

With this background, and to complement existing models by a preferably simple one, the Model-Data-Integration and Global Diagnostic Modelling groups at the MPI-BGC are currently developing a global model-data fusion framework for the coupled carbon and water cycle on land. Under the working title ‘SINDBAD’ different methods to combine heterogeneous, up-to-date earth observation data streams and biogeochemical models with low complexity are investigated. On the one hand, the models shall reproduce observed hydrological and biogeochemical patterns reasonable, and on the other hand serve to understand the influence and interactions between the different data streams. Therefor the models shall only consider components of the Earth’s system that are essential for consistency with observational data. Besides, applying the simplest model structure feasible is desirable in order to reduce problems of overparametrization and parameter equifinality, as well as to keep the model’s runtime as short as possible.

Combining diverse, spatially and temporally resolved data in models has proven to be an essential tool to estimate biogeochemical and hydrological processes, storages and associated changes (Döll et al., 2015). Among such data, terrestrial water storage anomalies obtained by the twin-satellites of the Gravity Recovery and Climate Experiment (GRACE) mission recently showed high potential for application in hydrological modelling (Döll et al., 2015; Güntner, 2008; Werth et al., 2009). GRACE provides continually measurements of the Earth's gravitational field, and after correction for atmospheric and oceanic effects, variations in this signal mainly represent changes in the terrestrial water storage (TWS) (Xie et al., 2012). As TWS integrates water storages in the form of snow, ice, surface, soil and ground water, it potentially allows inferences about spatial and temporal variations of its components. This is especially interesting in terms of a coupled carbon hydrologic model, since water availability, in particular in the soil, determines the distribution and productivity of plants, and thus represents one of the main links between both cycles (Gerten et al., 2004). Due to its global coverage and since in contrast to other satellite-based products it does not rely on surface conditions, GRACE TWS is a unique data source to quantify spatio-temporal variations of the Earth's water storages (Güntner, 2008). Thus, GRACE estimates are recommended to be used as model constraints, and in particular may be efficient in high latitudes where in situ hydrological measurements are sparse (Alkama et al., 2010; Werth et al., 2009). High latitudes additionally are in focus of research, as they are a potential carbon sinks and impacted by climate change more rapidly and significantly than other regions (Tallaksen et al., 2015).
Accordingly, GRACE TWS observations seem promising to combine in the SINDBAD framework and constrain its hydrological compartments.
In this context, this master thesis focuses on modelling large-scale hydrologic dynamics by developing a simple model that is consistent with spatial and temporal GRACE TWS patterns and allows to draw inferences about variations in soil moisture resp. plant available water in high latitudes.

1.1 State of the Art

At the present moment, a variety of models exist that are capable to simulate hydrological processes from regional to global scales. Thereby the objectives, applied input data, model structures as well as methodologies to quantify water fluxes and storages differ largely (Döll et al., 2015; Haddeland et al., 2011; Müller-Schmied et al., 2014; Sood and Smakhtin, 2015; Teuling et al., 2010). To evaluate the models' hydrological performance and if required to estimate their parameter values by calibration, simulations are commonly compared to observed discharge of large river basins. These point-measurements are most suitable as they integrate over processes in the whole upstream basin and for long time represented the only measurement

of a hydrological variable with acceptable accuracy and coverage (Döll et al., 2015; Güntner, 2008). However, reasonable conformity in terms of discharge does not guarantee proper representation of all hydrological processes within the basin (Güntner, 2008). To review and improve the credibility of simulations, additional information on other hydrological variables and fluxes should be considered by performing multi-criteria calibration and validation (Sood and Smakhtin, 2015).

While in-situ measurements are sparse and do not capture spatial heterogeneity, increasing time series and enhancement in the quality of remote sensing products with global coverage facilitate a new source of observational data streams (Döll et al., 2015; Sood and Smakhtin, 2015). The benefits of such satellite-based products have already been demonstrated by various studies applying large-scale calibration and validation (Immerzeel and Droogers, 2008; Livneh and Lettenmaier, 2012; Parajka and Blöschl, 2008).

Since the start of the GRACE mission in 2002, research has especially focused on variations of terrestrial water storage and its incorporation into hydrologic modeling (Döll et al., 2015). GRACE TWS estimates have been analyzed using hydrologic models to estimate solid precipitation (Seo et al., 2010), snow mass (Niu et al., 2007), evapotranspiration (Ramillien et al., 2006; Rodell et al., 2004), discharge (Syed et al., 2005) and to gain insights in regional groundwater variations (Feng et al., 2013; Rodell et al., 2009).

Besides, simulated TWS has been validated against GRACE TWS data for large river basins in several studies (Alkama et al., 2010; Döll et al., 2014; Kim et al., 2009; Ngo-Duc et al., 2007; Schmidt et al., 2008; Swenson and Milly, 2006; Syed et al., 2009). In general, those studies revealed a qualitatively good agreement in the seasonal dynamics and continental-scale patterns of simulated and observed TWS. However, for some regions they found discrepancies in the phase and amplitudes of TWS variations, which besides to shortcomings in the meteorological forcing are commonly attributed to model structure and process representation. Thus, integration of GRACE in multi-criteria calibration could considerably improve large-scale hydrological modelling (Güntner, 2008; Ngo-Duc et al., 2007; Syed et al., 2009). This has been proven by Werth et al. (2009), who calibrated a global hydrological model against river discharge and GRACE TWS variations and by doing so improved the simulations regarding both constraints. Similarly, Xie et al. (2012) calibrated a semi-distributive model for basins in Sub-Saharan Africa and found that the model performs well in semi-arid and sub-humid areas. Further, they unveiled that model-based partitioning of TWS in its components may highly be uncertain and thus requires future efforts. This demand is also highlighted by Güntner (2008), who sees the separation of GRACE's integral mass variations into its components as "fundamental for hydrological applications". Therefore, they suggest to apply auxiliary in-situ and remote sensing based data to include information on individual storages.

Among the components of TWS, soil moisture plays a major role in terrestrial hydrology, yet modelling its dynamics remains a major challenge due to the strongly non-linear behavior and related feedbacks of these processes (Hagemann and Gates, 2003; Todini, 1996). Improvement and confidence of its spatial and temporal estimates thus is highly desirable. Milzow et al. (2011) achieved good results when combining satellite based surface soil moisture, altimetry and GRACE TWS data to calibrate a hydrologic model for the Okavango catchment in Southern Africa. Similar, large scale measurements achieved from the European Space Agencies' (ESA) Soil Moisture and Ocean Salinity (SMOS) and NASA's Soil Moisture Active Passive (SMAP) missions may be useful to constrain simulated soil moisture. However, they still have limited value as the microwave signals can only capture the upper 5 cm of soil (Döll et al., 2015; Lettenmaier et al., 2015).

In high latitudes, snow mass is found to be the primary component of TWS (Niu et al., 2007; Rangelova et al., 2007). The benefits of integrating satellite-based snow estimates as a modelling constraint have been shown for instance by Parajka and Blöschl (2008). They performed multi-criteria calibration using MODIS snow cover data for catchments in Austria. While runoff performance did not change significantly, the simulated snow dynamics improved as evaluation against ground-based measurements showed. However, MODIS as a multispectral sensor suffers from cloud coverage and provides information on snow extent instead of the amount of water contained within the snow pack. Passive microwave sensors in contrast allow the estimation of the snow water equivalent (SWE) while being unaffected by cloud coverage. On the downside, passive microwave estimates of SWE are known to have large errors for either wet, deep, or ice and hoar containing snow packs (Lettenmaier et al., 2015). However, recently assimilation products of passive microwave satellite data and ground observations as for instance provided by the ESA's GlobSnow project improved these inaccuracies and seem a promising constraint for modelling snow dynamics (Hancock et al., 2013; Liu et al., 2014).

1.2 Objective and Research Questions

In the context of current efforts at the MPI-BGC to establish the SINDBAD model-data fusion framework, this master thesis aims is to develop a simple, globally applicable model that simulates hydrological fluxes and pools. Therefor advantage shall be taken of pertinent, state-of-the-art earth observation based data to force and constrain the model. Among such data, TWS anomalies obtained by the GRACE mission recently showed high potential for application in hydrological modelling. As TWS comprises all water stored near the land surface, it especially seems promising to assess soil moisture dynamics more accurately, which are

difficult to model, yet represent the main link between the hydrological and carbon cycle. In these terms, it is especially interesting how the TWS is composed spatially and temporally, meaning what the relative contributions of its components to changes in the combined signal are. In high latitudes of the Northern Hemisphere, the TWS variations are assumed to be mainly influenced by snow accumulation and ablation. As the Northern Hemisphere due to its greater land area and amount of biosphere additionally plays a major role in the carbon cycle, yet in situ hydrological measurements are sparse, this thesis focuses on the dynamics of TWS variations in snow affected regions on the Northern Hemisphere.

The intention is to develop a model that simulates the water pools

a) snow pack (snow water equivalent),
b) soil moisture, and eventually
c) groundwater

as well as the water fluxes

a) snow accumulation
b) snow melt
c) sublimation, and
d) runoff

by forcing the model with observation based products of precipitation, radiation, air temperature, and evapotranspiration. As numerous approaches for modelling the hydrological cycle and its components exist, this study will implement and examine different of these model formulations to identify the most appropriate, with a minimum of effective parameters and the maximum consistency with observational data.

In contrast to the common aim of hydrologic models to simulate average (annual) water flow accurately, the primary interest of this thesis lies in a sufficient representation of available water. Thus, the latest GRACE TWS estimates will be used for calibration of the model's parameters rather than runoff data. Following Güntner (2008), additional information shall be included to gain confidence when separating TWS in its individual storage components. As TWS dynamics in high latitudes are mainly influenced by snow accumulation and melt, satellite-based observations of snow water equivalent (SWE) obtained by the GlobSnow project will be used. Thus, this thesis will follow the recommendation of Döll et al. (2015) and Sood and Smakhtin (2015) and apply multi-criteria calibration. Since satellite-based SWE estimates are known to be subject to high uncertainties for deep snow conditions (Lettenmaier et al., 2015), one major question is how to combine the distinct data streams in the calibration process to take advantage of their informative content, yet consider these known issues.

Summarized, the objectives of this master thesis are threefold. The first two are methodological issues, concerning 1) which model formulation is 'best' and 2) which optimization method is most suitable. Finally, the resulting 'best' optimized model shall be used to address the scientific questions 3) whether variations in TWS can be reproduced and how the TWS signal is composed.
In detail, this thesis aims to answer the following questions:

1) Which model formulation for soil moisture dynamics and runoff generation is most appropriate and consistent with earth observation data, in particular with GRACE TWS and GlobSnow SWE estimates? How well does the model reproduce observations? Does spatial varying of model parameters improve the performance considerably?

2) Which optimization method is most suitable for multi-criteria model calibration? What search algorithm provides the best results yet is the less time-consuming? How should the cost function by formulated to consider known data issues?

3) What are the relative contributions of snow pack, soil moisture, and groundwater dynamics on terrestrial water storage variations? How do they vary spatially and temporally? Can trends of the GRACE TWS signal be reproduced by the model?

Subsequent to this introduction, the master thesis is subdivided into six chapters. Chapter 2 provides an overview on relevant theory, including a short description of the hydrological cycle, a review on (global) hydrological modelling and common process representations, as well as model calibration and evaluation methods. Chapter 3 explains in detail the developed model variants, the used data products and the experimental design. Following to a brief characterization of the study area based on the applied data in chapter 4, chapter 5 presents the obtained results. They are discussed regarding the research questions in chapter 6 and chapter 7 finally summarizes the findings of this thesis.

2 Theory

The following chapter provides an overview on macro-scale hydrological modeling. After a brief description of the hydrological cycle, types of hydrological models and features of existing global hydrological models are expounded. Subsequently, common methods for model optimization and evaluation are presented.

2.1 The Hydrological Cycle

The hydrological cycle is a complex and dynamic system that is strongly interconnected with the energy and biogeochemical cycles (Hagemann, 2011; Pagano and Sorooshian, 2006). It describes the continuous movement and retention of water through and in the Earth's spheres, driven by solar energy and gravitation (Brooks et al., 2012). A general scheme of the hydrological cycle, its components and fluxes is depicted in Fig. 1. As it illustrates, major reservoirs as ice and snow, surface water, soil, groundwater, ocean and atmosphere are interconnected by physical processes such as precipitation, evaporation and runoff. These processes cover various spatial scales and are highly variable in time and space (Hagemann, 2011).

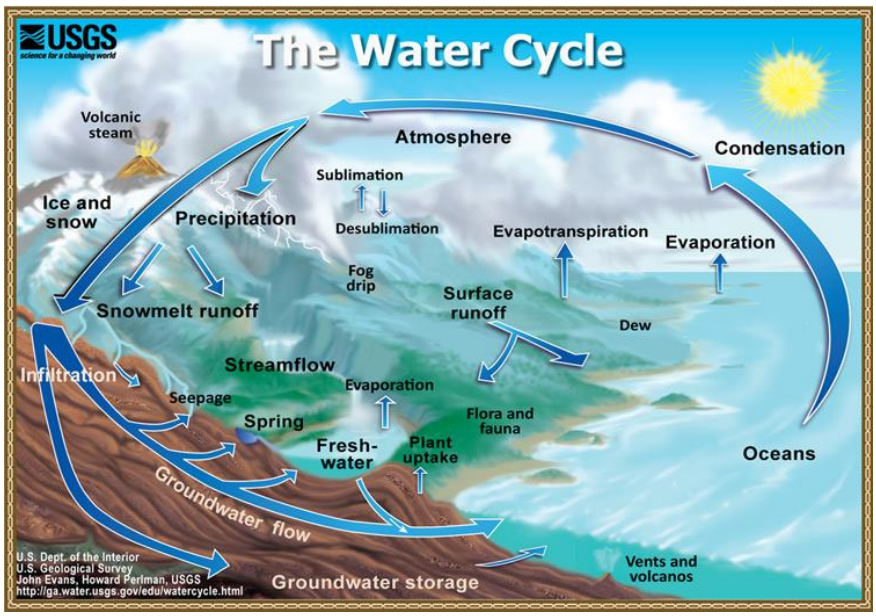


Fig. 1: Diagram of the natural water cycle (without human alterations) (Evans and Perlman, 2015).

In general, atmospheric water vapor precipitates on the Earth's surface, eventually flows as runoff to the ocean or inland water sinks while being transferred through the soil, the ground and/or surface water bodies, and finally evaporates again. Thereby, water fluxes and storage

conditions are strongly interconnected and influenced by various climatic and physio-geographic factors. For instance, dependent on temperature, precipitation most commonly occurs as rain or snow, but also includes drizzle, sleet, hail, and in a broader sense fog, dew and frost. Besides temperature, also wind, topography, vegetation and physical obstructions determine the deposition and accumulation of snow and ice. Whether snowmelt and liquid precipitation infiltrate depends on various factors such as the moisture status of the soil, its maximum water-holding capacity, the network and size of pores within the soil matrix, the condition of the soil surface including the vegetation cover, as well as rainfall and snow melt rate (Blume et al., 2010). Additionally, human activities influence the hydrological cycle among others by building reservoirs, withdrawal from water storages, or land-use activities that modify vegetation and water bodies, which in turn influences for instance evapotranspiration and the distribution of snow (Brooks et al., 2012).

While its allocation in storage or circulation varies in time, the mass of water remains constant on the global scale. Thus, the components of the hydrological cycle can be estimated for a distinct area by the water balance (Fig. 1). Accordingly, input by precipitation P equals the output represented by evapotranspiration ET (comprising transpiration, evaporation from interception, bare soil and open water surfaces) and streamflow Q (including surface runoff, interflow and base flow) and the change of storages ΔTWS such as ice, snow, soil moisture, ground water and surface water bodies (Schmidt et al., 2008).

$$P = ET + Q \pm \Delta TWS \tag{1}$$

2.2 Hydrological Models

To gain a better understanding of hydrologic phenomena and how these are affected by changes in climate and land use, the complex hydrological cycle can be represented in simplified terms by mathematical models. The following section provides an overview on different types of these hydrological models as well as their features regarding macro-scale application.

2.2.1 Types of Hydrological Models

Local to regional modeling of the hydrological cycle and its components has a long tradition in hydrological science and a variety of hydrological models (HM) have been developed. Numerous attempts to classify these HMs exist, where the most often used classification methods depend on the physical process description and the spatial discretization (Refsgaard, 1996; Xu, 2002).

According to their process representation, HMs are classified as empirical, physically based or conceptual. Empirical models use observation oriented, data driven methods and do not consider features and processes of the hydrological system. They apply statistical methods as regression and correlation or machine learning techniques to reveal the functional relationship between concurrent input and output data streams. Although they have high predictive power, these 'black box' models strongly depend on the used training data sets and allow little process-related conclusions (Devia et al., 2015). In comparison, physically-based models rely on fundamental physical laws. They use measurable state variables and their parameters have a physical interpretation. However, physical-based models as SWAT (Soil and Water Assessment Tool) (Neitsch et al., 2002), SHE (Systéme Hydrologique Européen) (Abbott et al., 1986) and tRIB (TIN-based Real-Time Basin Simulator) (Ivanov et al., 2004) need a large number of parameters to describe the physical characteristics of the catchment and suffer from scale related problems (Devia et al., 2015; Refsgaard, 1996).

Conceptual models as the HBV model (Hydrologiska Byråns Vattenbalansavdelning) (Bergström, 1976; Bergström, 1995; Lindström et al., 1997), TOPMODEL (Topographic Model) (Beven and Kirkby, 1979; Beven et al., 1995) and SAC-SMA (Sacramento Soil Moisture Accounting Model) (Burnash and Singh, 1995) are intermediate between empirical and physical-based models. Similar to the latter, they consider physical laws, but in highly simplified form. Using semi empirical equations, they describe processes that link a number of reservoirs, each representing a component of the hydrological cycle. Because of the simplified form, model parameters still have some physical meaning, yet cannot be assessed from field data alone but usually need calibration (Refsgaard, 1996; Xu, 2002).

In terms of spatial discretization, lumped models with no or only coarse spatial partitioning can be distinguished from semi-distributed and fully distributed models, that incorporate variability of the hydrological parameters, processes and inputs. Therefore, the latter divide the area of interest in elementary units and compute hydrological processes for each unit separately while allowing interaction between them. Frequent concepts for spatial distribution are raster grids (e.g. TOPMODEL), triangulated irregular networks (TINs, e.g. tRIBS) and areas of common characteristics and process dynamics as hydrologic response units (HRUs) (Flügel, 1995) delimited by polygons (e.g. J2000). Semi-distributed models partition the catchment into separate areas as sub-basins that are treated as homogenous within themselves and may not be explicitly spatially located (e.g. SWAT), whereas lumped models treat the catchment as a homogenous whole, with all parameters and variables representing average values of the total area (e.g. SAC-SMA).

In general, most conceptual models are lumped, whereas most physically-based can be considered distributed (Xu, 2002).

Traditionally, those stand-alone HMs are designed for and applied at the catchment- or basin-scale. They require measurements of precipitation, air temperature, global radiation or sunshine duration, relative humidity and wind for meteorological forcing and are usually calibrated against observed discharge at the respective gauge. To account for characteristics of the local environment, they need physio-geographic information as elevation, land cover/land use, soil porosity, field capacity and root depth and include a number of parameters that have to be estimated regionally (Sood and Smakhtin, 2015).

On the catchment-scale those HMs have been proven as essential tools in flood forecasting and regional water resource management (Gupta, 2011; Xu, 2002). However, since the late 1980s, issues of global climate change arouse the need for large-scale and global estimations of hydrological fluxes and storages (Xu, 2002). Initially, this has been impeded due to data scarcity at the global scale, as the number of ground based observations is limited, not all input information needed is available globally and if, it is associated with high uncertainty. Nevertheless, with significant progress in satellite-based data acquisition and enhanced computational resources, uninterrupted, large-scale and global information became available. As a result, a number of macro-scale hydrological models evolved, that are capable to simulate spatio-temporal patterns of global hydrological fluxes and storages.

In principle, macro-scale HMs are similar to traditional HMs, but are adapted to global process knowledge and available data streams. Thus, they differ in time scales and spatial resolution, the detail of process description, and parameter estimation approaches. The majority of macro-scale HMs can be categorized as conceptual models, as they are more frugal in data requirements than physically-based, but still capable to translate the effects of global change in the hydrological cycle appropriately. However, continuous development by adding more functionality (e.g. in terms of process description) and by the availability of finer global data sets led to a significant increase not only in the number, but also in the capabilities of macro-scale models. Hence, current models are more complex than earlier attempts and the conceptual process description became more physically based over time (Döll et al., 2015).

Among macro-scale models, literature often distinguishes between global hydrological models (GHM) and land surface models resp. land surface schemes (LSM) (Döll et al., 2015). While the latter evolved as land components of global circulation models (GCM), GHMs as WaterGAP (Döll et al., 2003) and WASMOD-M (Widén-Nilsson et al., 2007) are explicitly designed to assess the state of water resources by modelling the water balance and often take into account human water use.

In comparison, LSMs as BATS (Biosphere-Atmosphere Transfer Scheme of the U.S.'s National Center for Atmospheric Research's Whole Atmosphere Community Climate Model) (Dickinson, 1986) or H-TESSEL (land surface scheme of the European Centre for Medium-Range Weather Forecasts' operational weather forecast system) (Balsamo et al., 2009)

additionally solve the energy balance at high temporal resolution (hours), simulate evapotranspiration and snow melt in a less conceptual manner and represent soil with a higher vertical resolution (Haddeland et al., 2011). On the other hand, LSMs often lack representation of groundwater, surface water bodies, lateral flow routing and human alterations of the water cycle (Sood and Smakhtin, 2015).

In addition to GHMs and LSMs, dynamic global vegetation models (DGVM) as LPJ (Lund-Potsdam-Jena Model) (Gerten et al., 2004) incorporate hydrological processes to simulate the dynamics of ecosystem processes. Despite, as they focus on climate-vegetation interactions, they similar to LSMs usually do not represent lateral water movement and surface water bodies but include dynamic representation of the energy budget (Döll et al., 2015; Scheiter et al., 2013). However, due to similarities in particular areas between individual GHMs, LSMs and DGVMs, existing models can hardly be strictly classified and the grouping depends on considered aspects (Döll et al., 2015; Haddeland et al., 2011). Hence, in the following, the term 'GHM' will refer to macro-scale models in general.

2.2.2 Features of Global Hydrological Models

In general, GHMs adopt process formulations from catchment-scale HMs to estimate water flows on the Earth's land areas as evapotranspiration, river discharge and groundwater recharge, as well as water storage (variations) in different compartments (snow, soil, groundwater, surface water bodies) based on precipitation and other global meteorological data. GHMs partition precipitation into evapotranspiration, soil moisture and runoff while tracking the transfer of water through the different storage compartments. The main output for each model unit is streamflow, which is aggregated through the river network by routing methods to generate basin scale discharge (Döll et al., 2015; Sood and Smakhtin, 2015).

While this general scheme coincides, existing GHMs differ in their objectives, input data, spatio-temporal resolution, complexity and structural issues as the representation of anthropogenic influence, snow, vegetation, evapotranspiration, soil, groundwater and runoff processes as well as whether they account for sub-grid heterogeneity or not (Haddeland et al., 2011).

Tab. 1 summarizes a recent overview on existing GHMs and some of their features. In the following, the differences between the models will be addressed, with a special focus on divergent approaches of representing hydrological processes.

Tab. 1: Overview on existing global hydrological models,including their spatio-temporal resolution and structural characteristics (altered according to Haddeland et al. (2011) and Sood and Smakhtin (2015)).

Model		Resolution		Snow (Melt)	Potential Evapo-transpiration	Soil Layers / GW	Vegetation	Runoff	Human Altera-tion	References
		Spatial	Temporal/ Time Step							
WBM-WTM	Water Balance Model – Water Transport Model	0.5°	monthly / month	Eq. on T, P	Korzoun et al. (1978)	1 / -	3 types (forest, shrubland, grassland)	bucket/ linear	-	Vörösmarty et al. (1998)
WBM_{plus}	Water Balance Model plus Irrigation	0.5°	monthly / day	Eq. on T, P	Hamon (1963)	1 / yes	3 types (forest, shrubland, grassland) + crop types	bucket	yes	Wisser et al. (2010)
HDTM 1.0	Hydro Dynamic Model	0.5°	monthly / month	Eq. on T, P	Korzoun et al. (1978)	1 / yes	3 types (forest, shrubland, grassland)	bucket/ linear	yes	Wisser et al. (2010), Marchenko et al. (2008)
Macro-PDM	Macro Probability Model	0.5°	daily / day	dd	Penman, Penman-Monteith, Priestley-Taylor	1 / yes	2 types (forest, grassland)	beta	-	Arnell (1999) Gosling &Arnell 2010
MPI-HM	MPI Hydrology Model	0.5°	daily / day	dd	Thronthwaite, Input	1 / yes	-	bucket/ beta	-	Hagemann & Gates 2003, Hagemann & Dümenil 1998
GWAVA	Global Water Availability Assessment model	0.5° / 0.1°	monthly / day	dd	Penman-Monteith	1 / yes	4 types (trees, bushes, grass, bare soil)	bucket/ beta	yes	Meigh et al. (1999)
VIC	Variable Infiltration Capacity model	2°	daily / day	EB	Penman-Monteith	arbitrary, mostly 2-3	13 LC types + water	beta	-	Nijssen et al. (2001) Gao et al. (2010)
LaD	Land Dynamic model	1°	daily / hour	EB	EB	1 / yes	10 types	bucket/ linear	-	Milly and Shmakin (2002)

dd: day-degree approach; EB: equation based on energy balance; WB: equation based on water balance; LC: land cover; bucket: saturation excess runoff; beta: direct runoff as a nonlinear function of soil moisture; linear: temporal delay of runoff assuming a linear reservoir

Continutation of Tab. 1.

Model		Resolution Spatial	Resolution Temporal/ Time Step	Snow (Melt)	Potential Evapo-transpiration	Soil Layers / GW	Vegetation	Runoff	Human Altera-tion	References
WaterGAP	Water – Global Analysis and Prognosis model	0.5°	monthly / day	dd	Priestley-Taylor	1 / yes	16 LC types	beta	yes	Döll et al. (2003), Müller-Schmied et al. (2014)
PCR-GLOBWB	PCRaster Global Water Balance model	0.5°	daily / day	dd	Penman-Monteith	2 / yes	3 categories (natural, rainfed, irrigated), tall/short	beta	-	van Beek and Bierkens (2008)
LPJ/ LPJmL	Lund-Potsdam-Jena/ managed Land model	0.5°	daily / day	dd	Priestley-Taylor	2 / -	7 natural functional types, 12 crop functional types	bucket	yes	Bondeau et al. (2007), Gerten et al. (2004)
WASMOD-M	Water and Snow balance Modeling system	0.5°	monthly / month	Eq. on T	Eq. on T, relHum, Input	1 / -	-	linear	-	Widén-Nilsson et al. (2007)
H-08		1° / 0.5°	daily / 3 h	EB	EB	1 / -	crop growth model	bucket/ beta	yes	Hanasaki et al. (2008)
ISBA-TRIP	Interactions between Soil, Biosphere, and Atmosphere – Total Runoff Integrating Pathways	1°	daily / 20 min, 1 h	EB	WB of soil layer	3 / -	CORINE LC	beta	-	Alkama et al. (2010)

dd: day-degree approach; EB: equation based on energy balance; WB: equation based on water balance; LC: land cover; bucket: saturation excess runoff; beta: direct runoff as a nonlinear function of soil moisture; linear: temporal delay of runoff assuming a linear reservoir

Input Data and spatio-temporal Resolution

For model forcing, usually meteorological data derived from satellite products, provided by the Climate Research Unit (CRU) or daily reanalysis data as the European Centre for Medium-Range Weather Forecasts (ECMWF) ERA datasets are used.

Due to the availability of these global datasets, most GHMs have a spatial resolution of 0.5° or 1° and equivalent to their input, are based on the grid format (Sood and Smakhtin, 2015). Recently, data with finer resolution along with faster computational devices have facilitated the development of GHM versions with higher resolution (e.g. a 5' version of WaterGAP) (Döll et al., 2015). Although this trend is promoted by projects as the HyperHydro initiative, which aims to perform global-scale hydrological modelling on grid cells of 100 – 1000 m (Wood et al., 2011), hyperresolution GHMs are not yet common.

The modelled time steps range from one month to less than one day, with the majority of GHMs simulating in one-day resolution. If nevertheless monthly input data is used, statistical downscaling methods are applied (Sood and Smakhtin, 2015).

Structural differences of GHMs

Snow

Some GHMs explicitly account for snow accumulation and melt. They partition precipitation into rainfall and snowfall either using a threshold temperature, which usually is 0 °C or 1 °C, or into a combination of rain and snow between an upper and a lower threshold temperature (Haddeland et al., 2011). For instance, in WaterGAP all precipitation falls as snow if the temperature is below 0 °C, whereas the MPI-HM simulates a combination of rain and snow between threshold temperatures of -1.1 °C and 3.3 °C (Hagemann and Gates, 2003). Apart from these traditional approaches, rainfall and snowfall can be distinguished based on satellite observations of snow depth. This method is applied in the Global Land Evaporation Amsterdam Model (GLEAM), that considers precipitation as snowfall when snow depth is more than 10 mm (Miralles et al., 2011).

Snow melt can be simulated based on energy balance or, in a more conceptual manner, with the degree-day approach. The latter is a temperature-index approach, which assumes a relationship between ablation and air temperature and calculates snow melt by multiplying the difference between daily temperature and a melt threshold temperature with the degree-day factor (He et al., 2014). A common used threshold temperature is 0 °C (Müller-Schmied et al., 2014). MPI-HM and PCR-GLOBWB use the day-degree formula according to the HBV model, applying a degree-day factor of 3.22 mm $°C^{-1}$ (Stacke et al., 2011). In WaterGAP 2.0 the degree-day factor is land cover specific, with 2 mm $°C^{-1}$ for forest and 4 mm $°C^{-1}$ for other land areas (Müller-Schmied et al., 2014). Other GHMs such as WBM and WBM_{plus} additional include the effects

of rain by calculating snowmelt using an empirical function of T and P if temperature exceeds the melt threshold of 1 °C (Vörösmarty et al., 1989; Wisser et al., 2010). Kustas et al. (1994) as well as Rango and Martinec (1995) suggest to include a radiation component into the degree-day approach in order to improve simulated melt rates (restricted day-degree approach).

Besides differences in the threshold temperatures and degree-day factors, another structural difference among GHMs is whether melted snow percolates through the snow directly into the soil or can be retained in the snow pack (Haddeland et al., 2011). For example, the MPI-HM and PCR-GLOBWB allow 6% of snow water equivalent to be retained as liquid water in the snowpack (Hagemann and Gates, 2003).

In addition to snow melt, sublimation decreases the snow pack potentially. This process can be estimated by adapting potential evaporation equations for ice. For instance, GLEAM and WaterGAP 2.0 both obtain sublimation over ice and snow covered areas by applying an adapted latent heat in the Priestley and Taylor evaporation equation (Miralles et al., 2011; Müller-Schmied et al., 2014).

Sub-grid variability is incorporated by some GHMs by delineating the snow covered fraction of the grid cell. One possibility to implement this, is to assume snow cover fraction to be linearly increasing between zero snow water equivalent and a threshold, above which the whole grid cell is assumed to be snow covered. For example, H-TESSEL defines this threshold at 15 mm snow water equivalent (Balsamo et al., 2009; ECMWF, 2014). In WaterGAP snow accumulation and melt is modeled for a 3’’ sub-grid, with grid temperature being modified to the sub-grids based on elevation information and an adiabatic lapse rate of 0.6 °C per 100 m (Müller-Schmied et al., 2014).

Vegetation

GHMs largely differ in the level of detail given to vegetation description and processes. While some, as H-08 (Hanasaki et al., 2008) and LPJmL (Bondeau et al., 2007), include an explicit crop growth model that can deal with the impact of CO_2 and temperature changes on plant physiology, others apply a land cover/land use classification of varying detail and define vegetation specific parameters as root depth and albedo (Sood and Smakhtin, 2015).

Evapotranspiration

Regarding evapotranspiration, GHMs use different methods for calculation and vary in the components they consider. Common methods to derive evapotranspiration are the Thornthwaite (1948), Penman (1948), Penman-Monteith (Monteith, 1965) and Priestley-Taylor (Priestley and Taylor, 1972) formula. While the Thronthwaite formula is based on the empirical relation between potential evapotranspiration and daily mean temperature scaled by the relative day length, the Penman approach is semi-empirical. It applies a number of meteorological observations as radiation, vapor deficit and ventilation (wind) and thus can be adapted to local

site conditions. Both, Priestley-Taylor and Penman-Monteith represent modifications of the Penman formula. The latter additionally includes two vegetation specific evaporation resistances, whereas Priestley-Taylor is based on less extensive measurements (Climate Service Center Germany, 2012). Besides, evapotranspiration can be calculated based on the energy balance within the model (H-08) or imported from an external data set (MPI-HM, WASMOD-M).

For instance, MPI-HM either calculates potential evaporation according to Thornthwaite or applies a concurrent input data set. The model considers interception evaporation from a canopy storage with a distinct capacity and directly distinguishes evaporation from bare soil from vegetation transpiration. Evaporation from bare soil is calculated as a linear function of the ratio of potential evapotranspiration and available soil moisture, while transpiration is set to maximum for soil moistures above a critical threshold and linearly decreases towards the wilting point. To account for sub-grid heterogeneity, evaporation and transpiration are further scaled to the vegetation fraction of the grid, which has to be provided as an input (Stacke et al., 2011).

In WaterGAP potential evapotranspiration is derived according to Priestley-Taylor. The model also includes interception evaporation, which is calculated following Deardorff (1978) as function of potential evapotranspiration and maximum canopy capacity depending on the seasonal-varying leaf area index (LAI) of the corresponding land cover. Actual evapotranspiration from soil is derived from the remaining potential evapotranspiration and the available water content in the land cover specific effective root zone compared to the maximum water capacity of this zone. For open inland waters the actual evapotranspiration is assumed to be equal to potential evapotranspiration (Döll et al., 2003; Müller-Schmied et al., 2014).

Soil

In general, some GHMs consider a single soil storage, others multiple soil layers (Sood and Smakhtin, 2015), between which capillary rise may be reproduced (PCR-GLOBWB). Further, the soil layer may have a fixed depth or depend on a vegetation specific root depth zone (WaterGAP, Zhang et al. (2008)). Some GHMs, as the HDTM 1.0 are capable to simulate permafrost dynamics and take into account frozen soil properties (Marchenko et al., 2008; Wisser et al., 2010).

For separation of liquid precipitation resp. throughfall from the canopy layer and snow melt into infiltrating water and surface runoff GHMs apply different conceptual methods.

An often applied approach is the bucket-type scheme first introduced by Manabe (1969). It assumes each grid cell to be a single soil water reservoir, which represents field capacity and thus water storage capacity of the soil. After a precipitation event, (direct) runoff will only occur if the reservoir is fully saturated (Hagemann and Gates, 2003).

Common modifications of the bucket scheme have been developed in the HBV model and the Arno model (Dümenil and Todini, 1992) to allow runoff generation even if the grid cell is not completely saturated. Thereto, the HBV approach according to Bergström (1995) describes runoff as a non-linear function of effective precipitation (sum of rainfall and snowmelt), soil moisture, maximum storage capacity, and a runoff coefficient, which determines the amount of effective precipitation converted into runoff (Alcamo et al., 2003). As this method is simple and independent of scale, it is widely applied, for instance in WaterGAP and simple water balance model (KM12) by Koster and P. Mahanama (2012).

The effect of increasing surface runoff resp. decreasing infiltration with increasing soil moisture is also achieved by representing the sub-grid variability in soil water capacity as it is done in the Arno scheme. This method assumes a statistical distribution of storage capacities within the model entity. The parabolic distribution function, whose form is defined by a parameter similar to the Bergström approach, describes the fractional area that is saturated and thus contributes to runoff generation. Many GCMs as well as the GHM VIC, Macro-PDM and PCR-GLOBWB apply this approach to determine infiltration and surface runoff. Other GHMs as the MPI-HM use an improved Arno scheme, which derives fractional saturation curves for each grid cell from land surface parameters instead of assuming a statistical distribution (Hagemann and Gates, 2003).

Another approach to estimate the partitioning of precipitation into soil moisture and runoff is shown by Zhang et al. (2008), who created a lumped conceptual model based on the Budyko demand-supply framework for modelling mean annual evapotranspiration (Budyko, 1961). Assuming rainfall to be the supply limit and available storage capacity in the root zone to be the demand limit, they delineate runoff from the demand/supply index and a retention efficiency.

Other, simple GHMs as WASMOD-M and WBM do not distinguish between different water storage compartments and derive the amount of total moisture per grid cell by calculating the water balance using estimated evapotranspiration and runoff (Widén-Nilsson et al., 2007).

Another very simplified approach originating from recession analysis assumes discharge to be proportional to a power of the catchments' total water storage (Brutsaert and Nieber, 1977). While this method is widely used in studying groundwater systems, Kirchner (2009) applied it in a simple dynamical system approach, which yields good results for catchments under humid conditions (Kirchner, 2009; Melsen et al., 2014; Teuling et al., 2010).

Runoff Generation and Groundwater

GHMs distinguish between different runoff components and incorporate water retention in differing ways. Further, they may include routing of the grid's total runoff through the stream network in order to compare modelled stream flow with discharge measurements.

While simple modeling approaches as the KM12 neglect water retardation, most GHMs incorporate a delayed runoff component by explicitly simulating a groundwater storage. Water is allocated to this storage through percolation either from the soil storage or as a fraction of surface runoff. LaD and MPI-HM for example model percolation from soil water storage depending on soil moisture (Stacke et al., 2011). Similar, VIC simulates drainage between soil layers as a function of the degree of saturation (Nijssen et al., 2001).

Other models, such as WBM_{plus} and WaterGAP, don't consider drainage from soil water storage and reduce soil moisture solely by evapotranspiration. Instead, WBM_{plus} allocates a predefined fraction of surface runoff to groundwater (50 %), while WaterGAP applies a heuristic approach based on slope characteristics and other physio-geographic data to partition surface runoff into direct runoff and groundwater recharge, depending on a soil texture specific maximum recharge rate (Müller-Schmied et al., 2014).

The groundwater itself is mostly represented as a linear storage compartment that releases base flow into surface water bodies or total runoff. Thereby, the outflow usually is proportional to the storage content and calculated by dividing the reservoir content by the retention time of the reservoir, which is defined as average residence time of water within the reservoir (Hagemann and Dümenil, 1997). Those retention parameters are for example 0.01 d^{-1} (WaterGAP) and 0.0167 d^{-1} (WBM_{plus}).

In WaterGAP the sum of base flow and direct runoff equate to total runoff of each cell, which is further led through a series of storages representing wetlands, lakes and reservoirs until reaching the river segment. Subsequently, streamflow is routed through the stream network (Müller-Schmied et al., 2014). Similar to WaterGAP, VIC distinguish between direct runoff without delay and base flow, which originates from the deepest soil layer (Nijssen et al., 2001). Although it does not incorporate a distinct groundwater storage, H-08 distinguishes between direct runoff occurring when soil water storage is saturated, and a constant runoff component as exponential function of soil water storage (Hanasaki et al., 2008).

Similar to H-08, WASMOD-M includes no groundwater storage, yet describes a fast and slow runoff component, but uses linear reservoirs with two different retardation parameters. Thereby slow runoff solely originates from the water storage, while fast runoff emerges additionally from rainfall and snow melt. Both sum up to total runoff, which is then routed through the stream network without any delay (Widén-Nilsson et al., 2007).

In MPI-HM, both direct runoff and base flow are modelled as reservoirs with specific retention times, too, but retention is based on the slope of the grid cell. It further represents river flow within the grid cell as a linear flow cascade with five reservoirs (Hagemann and Gates, 2003).

As KM12 considers only one runoff component, Orth et al. (2013) found it necessary to impose a delay parameter that accounts for retardation of water transport in the catchment in order to make modeled runoff comparable to streamflow measured at a stream gauge site.

However, gradient based groundwater flow between grid cells usually is not represented in GHMs due to the extreme lack of information on groundwater distribution and table on global-scale (Döll et al., 2015).

Anthropogenic Influence

Human alterations of the hydrological system are incorporated in GHMs as WaterGAP, and H-08. For instance, WaterGAP includes a separate sub-model, the Global Water Use Model, that takes into account basic socio-economic factors that lead to domestic, industrial and agricultural water use, which then is subtracted from the water availability simulated by the Global Hydrology Model (Müller-Schmied et al., 2014).

2.3 Model Calibration and Evaluation Methods

In HMs, especially conceptual ones, parameters do not always correspond to physically measureable properties and thus cannot be determined directly. Therefore, they have to be estimated indirectly, by searching a parameter combination that leads to an optimal match between modelled variables and concurrent observations. This process of calibration can either be conducted manual, by 'trial and error', or by using automatic optimization techniques (Fischer, 2013). While manual parameter adjustment is time- and cost-consuming and involves subjective judgement, automatic techniques based on mathematical algorithms are capable to solve complex problems fast and objectively (Xu, 2002). Often, calibration is impeded by parameter interactions and multiple parameter sets gain equal acceptable model performance (Beven, 2006). Due to this problem of equifinality, automatic techniques may generate unrealistic solutions and thus still require user expertise to verify the optimized parameter values. Additionally, parameters are usually adjusted within a range of possible values, that is based on knowledge of the hydrology in the respective area, literature values or measurements (Fischer, 2013; Xu, 2002).

Basin-scale HMs are traditionally calibrated against observed river discharge, as these point-measurements integrate over processes in the whole upstream basin (Döll et al., 2015). In contrast, GHMs have rarely been calibrated, with a few exceptions such as WaterGAP and WASMOD-M. For them, one or two parameters are adjusted by evaluation of simulated against observed river discharge. Calibration in a broader sense also includes tuning of adjustment factors in order to reduce the bias between model output and observations, which for example is performed for WaterGAP and WBM (Döll et al., 2015; Sood and Smakhtin, 2015).

The calibrated model's performance and suitability is evaluated in a validation process, by comparing the model's output with independent information, either from data of the same catchment not utilized for calibration (split sample test) or data from another basin (proxy basin test) (Klemeš, 1986). While simple split sample testing examines the transferability of the

model under stationary conditions, differential split sample testing checks for the model's validity under transient conditions by dividing the time series e.g. in a wet and dry period. Similar, the proxy basin test, which verifies spatial transferability, can be combined with differential split sample testing (proxy basin differential split sample test) (Xu, 2002).

To assess model performance objectively during both, calibration and validation, statistics that estimate the error between simulated and observed variables are used (Legates and McCabe, 1999). In the course of this, errors of the observational data are usually neglected and measurements are assumed to represent the ground-truth (Döll et al., 2015; Moriasi et al., 2007). In general, GHMs are calibrated and evaluated either on the ecoregion, climate or continental scale, or, similar to traditional HMs, against discharge observations from large river basins, whereby the number of considered basins varies depending on the study (Sood and Smakhtin, 2015). Although most GHMs run on daily time steps, calibration and evaluation is conducted on monthly or long-term annual scales.

The following sections introduce some frequently used efficiency criteria and explain automatic optimization techniques more in detail.

2.3.1 Model Efficiency Criteria

Model calibration and evaluation utilizes efficiency criteria as a quantitative measure of the conformability between simulated and observed variables. According to Moriasi et al. (2007) at least one absolute error index, one dimensionless statistic and one graphical technique, such as visual comparison for a first overview on model performance, should be evaluated.

Mean Square Error MSE

The Mean Square Error (MSE) is a commonly used error index that describes the mean difference between observations and model simulations (Eq.(2)). It indicates the error in squared units of the constituent of interest and ranges from 0 to ∞. As a value of zero indicates perfect fit, modelling aims to generate the least mean square error possible (Moriasi et al., 2007).

$$MSE = \frac{1}{N}\sum_{i=1}^{N}\left(x_{i_{obs}} - x_{i_{mod}}\right)^2 \qquad (2)$$

where x_{obs} = observed variable

x_{mod} = modeled variable

N = number of data points

Since, unlike the theoretical assumption, observation streams are subject to errors, this uncertainty should be considered in evaluation statistics (Döll et al., 2015; Moriasi et al., 2007). Thereto, the residuals $x_{obs} - x_{mod}$ are weighted with the uncertainty σ of the observations (Foglia et al., 2009). Accordingly, the weighted mean square error (wMSE) is calculated as:

$$wMSE = \frac{1}{N} \sum_{i=1}^{N} \frac{\left(x_{i_{obs}} - x_{i_{mod}}\right)^2}{\sigma_i^2} \tag{3}$$

where σ = uncertainty of x_{obs}

Root Mean Square Error RMSE

The Root Mean Square Error (RMSE) is the root of the MSE and similarly expresses the error between modelled and observed data, yet in the units of the constituent of interest. A value of zero indicates perfect fit regarding the modelled volume, while a RMSE less than half the standard deviation of the observed data may be considered low. As MSE, the criterion ranges from 0 to ∞ (Moriasi et al., 2007).

$$RMSE = \sqrt{MSE} \tag{4}$$

Percentage Bias PBIAS

The percentage bias (PBIAS) measures the average tendency of the modeled data to be larger or smaller than the observation. Similar to RMSE, the optimal value is zero. However, as PBIAS is expressed in percentage rather than the units of the constituent of interest, it allows to compare areas and seasons with widely differing values. It ranges from -100 to 100 [%], with positive values indicating model underestimation whereas negative values imply model overestimation bias (Moriasi et al., 2007).

$$PBIAS = \frac{\sum_{i=1}^{N}\left(x_{i_{obs}} - x_{i_{mod}}\right)\ 100}{\sum_{i=1}^{N} x_{i_{obs}}} \tag{5}$$

Pearson's correlation coefficient r and coefficient of determination R^2

The coefficient of determination R^2 is the square of the Pearson's product-moment correlation coefficient and describes the proportion of total variance in the observed data that can be explained by the model. As r, it is an index of the linear relationship between the data streams and is insensitive to additive and proportional differences. Further, both statistics are

oversensitive to outliers. The value range of r is -1 to 1, with high negative resp. positive values indicating a strong negative resp. positive relation (Legates and McCabe, 1999). R^2 accordingly ranges from 0 to 1, and in general values > 0.5 are considered acceptable (Moriasi et al., 2007).

$$R^2 = \left\{ \frac{\sum_{i=1}^{N} \left(x_{i_{obs}} - \mu_{obs}\right)\left(x_{i_{mod}} - \mu_{mod}\right)}{\left[\sum_{i=1}^{N} \left(x_{i_{obs}} - \mu_{obs}\right)^2\right]^{0.5} \left[\sum_{i=1}^{N} \left(x_{i_{mod}} - \mu_{mod}\right)^2\right]^{0.5}} \right\}^2 \qquad (6)$$

where μ_{mod} = mean modeled variable

μ_{obs} = mean of observed variable

Spearman's rho ρ

Spearman's ρ is defined as the Pearson correlation coefficient r between the ranked variables and, similar to r, it is insensitive to additive and proportional differences. In comparison to r, it has the advantage of not requiring a linear relationship between the data and is less sensitive to outliers. Thus, on the one hand, this statistic provides a more robust characterization of the correlation. On the other hand, it is associated with loss of information due to the data conversion into ordinal form. Spearman's ρ ranges from -1 to 1, with -1 and 1 indicating a monotone negative resp. positive relation between the variables, while a value of 0 implies no relation at all (Legates and McCabe, 1999).

Nash-Sutcliff Model Efficiency MEF

The Nash-Sutcliff Model Efficiency (MEF) (Nash and Sutcliffe, 1970) represents the criteria most widely used to evaluate the performance of hydrological models, as it is one of the best objective functions for reflecting the overall fit of a hydrograph (Moriasi et al., 2007). MEF determines the relative magnitude of the residual variance compared to the variance in the observed data. In comparison to r, it designates a good model fit only if the long-term average of the observational data stream is captured well (Döll et al., 2003). Its values range from -∞ to 1, with higher values indicating better agreement. Generally, values between 0 to 1 are acceptable, whereas values < 0 suggest that the mean of the observed data is a better predictor than the simulated value. However, due to applying the squared differences, MEF is oversensitive to extreme values (Legates and McCabe, 1999).

$$MEF = 1 - \frac{\sum_{i=1}^{N} \left(x_{i_{obs}} - x_{i_{mod}}\right)^2}{\sum_{i=1}^{N} \left(x_{i_{obs}} - \mu_{mod}\right)^2} \qquad (7)$$

Similar to wMSE, the Nash-Sutcliffe Model Efficiency can be weighted to account for the uncertainty of the observed variable:

$$wMEF = 1 - \frac{\sum_{i=1}^{N} \frac{\left(x_{i_{obs}} - x_{i_{mod}}\right)^2}{\sigma_i^2}}{\sum_{i=1}^{N} \frac{\left(x_{i_{obs}} - \mu_{mod}\right)^2}{\sigma_i^2}} \tag{8}$$

Kling Gupta Efficiency KGE

The Kling Gupta Efficiency (KGE) (Gupta et al., 2009) is a criterion inferred from the decomposed MEF. According to previous studies MEF consists of three distinct components representing the correlation, the bias and a measure of relative variability in the simulated and observed values (Gupta et al., 2009; Murphy, 1988; Węglarczyk, 1998). Because of inherent systematic problems as the interplay of these components when optimizing MEF, Gupta et al. (2009) suggest calibration from the multi-objective view by focusing on the components as separate criteria. They propose KGE as the difference between unity and the Euclidian distance (ED) from the ideal point in the three-dimensional criteria space:

$$KGE = 1 - ED \tag{9}$$

$$ED = \sqrt{(r-1)^2 + (\alpha - 1)^2 + (\beta - 1)^2} \tag{10}$$

with r being Pearson's correlation coefficient, α being the variability error (11), and β representing the bias error (12).

$$\alpha = \frac{\sigma_{mod}}{\sigma_{obs}} \tag{11}$$

$$\beta = \frac{\mu_{mod}}{\mu_{obs}} \tag{12}$$

where σ_{mod} = standard deviation of modeled variables

σ_{obs} = standard deviation of observed variables

All three components are dimensionless and have their ideal value at unity. Hence, similar to MEF, KGE ranges from -∞ to 1, with unity being the optimal value. KGE also has a tendency to underestimate high observed values, but not as severe as MEF (Gupta et al., 2009).

Modification of Efficiency Criteria

By inspecting squared residuals, most efficiency criteria put emphasis on large errors (MSE, wMSE, RMSE, r, R^2, MEF, wMEF). Since errors of extremes tend to be larger than those associated with low values, they are oversensitive to high values and thus calibration often leads to fitting of extreme events (Krause et al., 2005). To overcome this oversensitivity, several approaches exist, including modifications of efficiency criteria and trimming of the data streams. The latter simply removes for example the 5 % of data points with the highest error (Trishchenko, 2002).

One often applied modification of MEF is the calculation with logarithmic values of x_{obs} and x_{mod}. Since logarithmic transformation flattens extreme values, the influence of low values is increased in comparison to MEF and thus the modified form is more sensitive to systematic model over- or underestimation (Krause et al., 2005). Instead of logarithmic transformation, Krause et al. (2005) suggest as another modification to consider in general the modulus of residuals $|x_{obs} - x_{mod}|$ rather than the squared form $(x_{obs} - x_{mod})^2$. This reduces significantly the overestimation of extremes and thus results in a better overall estimation.

2.3.2 Automatic Optimization

Automatic calibration involves the computation of the prediction error using an equation (cost function/objective function), which is usually based on one or more efficiency criteria, and an automatic optimization procedure (search algorithm) to search for parameter values that minimize the value of the cost function (Moriasi et al., 2007; Xu, 2002).

In general, automatic optimization procedures are mathematical search algorithms, that attempt to find the minima (or maxima) of a function $f(x_1 \ldots x_n)$ by iteratively modifying the vector $X = [x_1 \ldots x_n]$ within preset constraints, that define the subset of natural numbers in which X is permitted to take values (Soliman and Mantawy, 2011). In the terms of model calibration, $f(x_1 \ldots x_n)$ conforms to the cost function and X to the parameter set, which may be constrained by an upper and lower bound of reasonable parameter values or parameter interdependencies.

Starting from an initial point defined by the parameter set X_0, the algorithm performs intermediate calculations that lead to a new point X_1 associated with a lower cost function value. By repeating, successive approximation of the minimum is obtained. Finally, the parameter set associated with the least cost function value is regarded as the optimal parameter set (Soliman and Mantawy, 2011).

Overview on Optimization Methods

For hydrological models, $f(x_1 \ldots x_n)$ usually is a non-linear function of the parameter values. Such non-linear problems can be solved either classical by iterative methods that converge to a

solution, or by heuristic approaches that attempt to find an approximate solution based on concepts found in nature (Winker and Gilli, 2004).

In classical methods, the intermediate calculations involve evaluation of the cost function and possible constraints as well as estimations of the step size and direction in the function response surface in which to move. This estimation can be achieved by line search techniques and trust region methods. Further, one can distinguish between direct search methods, solely based on the differences in cost function values between the current point and adjacent points, and gradient-based methods that use the derivatives of the cost function. Among gradient-based approaches, methods as the Steepest Descent solely consider the gradient (first order partial derivatives), while others as the Newton method additional use the Hessian matrix (matrix of second order partial derivatives) or as Quasi-Newton methods an approximation of the Hessian. Regarding direct search, the minimum is characterized by negative differences to all adjacent points. For gradient-based methods the minimum is defined by a gradient of zero and a positive (semi-) definite (approximation of the) Hessian (Xu, 2002).
As it is virtually impossible to know when this point is reached, some termination criteria are required to determine when to stop the search. Those criteria refer to function convergence, when the value of the cost function over one or more iterations is not improved, parameter convergence, when the algorithm is not able to appreciably change parameter values and simultaneously improve the cost function, or, if no convergence criteria is met first, a maximum number of iterations (Xu, 2002).
A considerably disadvantage of this classical methods is, that they always converge in a local minimum, meaning a point where the cost function value is smaller than at nearby points, but not necessarily smaller than at all other feasible points as it is at a global minimum. As illustrated in Fig. 2, whether a local or the global minimum is obtained depends on the complexity of the response surface and the starting point. The set of starting points that will lead to the same local minima often is referred to as basin of attraction (MathWorks, 2015b).

Non-linear hydrological models usually have multi-modal response surfaces where several parameter combinations result in local minima of the cost function (Xu, 2002). To overcome this, global optimization techniques consider the entire parameter space during the search process. Thus, they need in general more function evaluations and more computational time compared to local methods (Fischer, 2013).

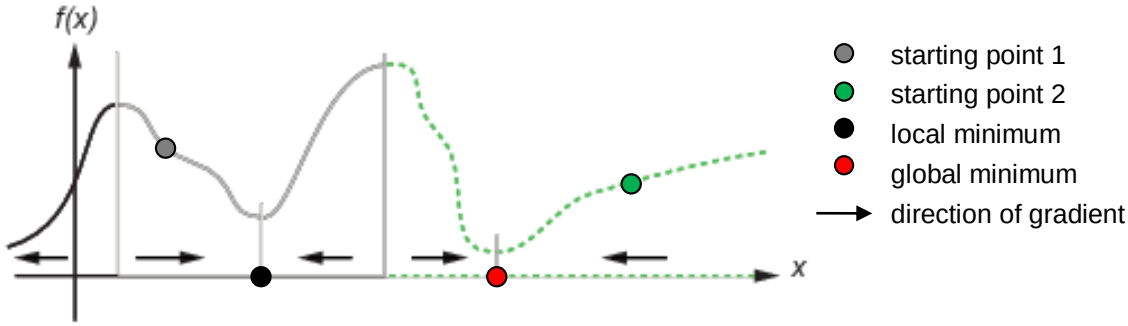


Fig. 2: Response surface of a one-dimensional cost function f(x) showing a local and a global minimum. Following the direction of the steepest gradient, starting point 1, as all other possible starting points of the grey line basin of attraction, will lead to the black local minimum, whereas starting point 2 and points of the basin of attraction depicted in green will result in the red global minimum (modified according to MathWorks (2015b)).

On the one hand, global techniques apply random search methods to define multiple starting points for local optimization and compare the local solutions to derive the global minima (MathWorks, 2015a). On the other hand, heuristic approaches as methods of particle swarm, simulated annealing, evolution strategy and genetic algorithms can be considered as global techniques (Fischer, 2013). Similar to classical optimization algorithm, heuristic approaches iteratively evaluate the cost function, but assess the new point in the parameter space by some generation rule. For example, evolutionary algorithms mimick processes of natural variation and selection. As these approaches don't make any assumptions of the response surface (e.g. don't require the cost function to be continuously differentiable), they are less likely to stop at local minima. Heuristic approaches don't guarantee to give the optimal solution, but significantly reduce high computation times needed by classical methods to converge if the resolve surface is highly complex (Winker and Gilli, 2004).

Lsqnonlin

The majority of calibration schemes in hydrology are related to classical least squares methods that seek parameter sets that minimize the squared residuals (Xu, 2002). One such technique is the Lsqnonlin algorithm within MATLAB's Optimization Toolbox.

Lsqnonlin is a non-linear least-square solver that aids to minimize the squared residuals of a non-linear function. The search direction within this gradient-based approach can be defined with the Levenberg-Marquardt algorithm (Marquardt, 1963). This method is based on the Gauss-Newton method, a simplification of the Newton method, which at each iteration substitutes the cost function by a squared approximation and attempts to minimize the residuals using least-squares. In comparison to the Newton method, the Gauss-Newton does not require second derivatives. Since in the Gauss-Newton method the sum of squares may not decrease at every iteration, Levenberg-Marquardt additionally forces a decrease by using a search direction that is a cross between the Gauss-Newton direction and the steepest descent direction. Therefore, Levenberg-Marquardt is regarded to be more robust than the original Gauss-Newton

attempt (MathWorks, 2015b). However, Lsqnonlin applying Levenberg-Marquardt is not able to handle bound constraints, which may be important to obtain reasonable parameter values.

Fmincon

Another local, gradient-based search algorithm provided within MATLAB's Optimization Toolbox that overcomes this problem is Fmincon. This solver attempts to find the minimum of constrained non-linear multivariable functions using the interior-point algorithm, which is most suitable for comprehensive and complex problems (MathWorks, 2015b). The algorithm's procedure is shown in Fig. 3 and illustrated in Fig. 4. Instead of minimizing the original inequality constrained problem, interior-point solves an approximation problem with a sequence of easier solvable equality constrained problems. Thereto, it employs a barrier function, which includes slack variables of the constraints and which ensures that the objective function value goes to infinitive if the constraints are crossed. By weighting the barrier function with a weight μ, the feasible region of a minimum is restricted (Fig. 4 a). The algorithm minimizes the approximation function with fixed, initially high μ, by solving a quadratic sub-problem (Fig. 4 b). For minimization of the sub-problem either direct steps using a line search based on Newton's method, or, if direct steps cannot be performed (e.g. approximation locally convex, Hessian not positive), a conjugate gradient step using the trust region method is applied (Waltz et al., 2006). For direct search, the Hessian is approximated by the Broyden-Fletcher–Goldfarb-Shanno (BFGS) method (Broyden, 1970), and thus can be considered as a Quasi-Newton approach (MathWorks, 2015b). The trust-region represents the neighborhood around the current point, in which the quadratic sub-problem reflects the behavior of the approximation reasonably. The minimum within this trust-region is estimated by conjugate gradient steps, which similar to the steepest descent don't rely on second order derivatives, but use a-orthogonal directions and thus avoid repetitious steps when iteratively moving towards the minima. The procedure of minimizing the sub-problem is applied repeatedly for decreasing values of μ, starting at the minimum obtained by the preceding step respectively. Thus, as μ decreases to zero, the minimum of the approximation should approach the minimum of the original cost function (Byrd et al., 2000).

Global Multi Start

As the solution of local methods strongly depends on the starting value, global methods are more appropriate for calibration of hydrologic models (Gupta et al., 2005).

The simplest way to search the entire parameter space is to define multiple, randomly sampled starting points and apply a local method for each. This is done for example by the global multi start solver (GS) provided in MATLAB's Global Optimization Toolbox. The algorithm generates *n* start points, which are uniformly distributed within the user-defined bounds. The solver runs all start points that are feasible with respect to bounds and inequality constraints.

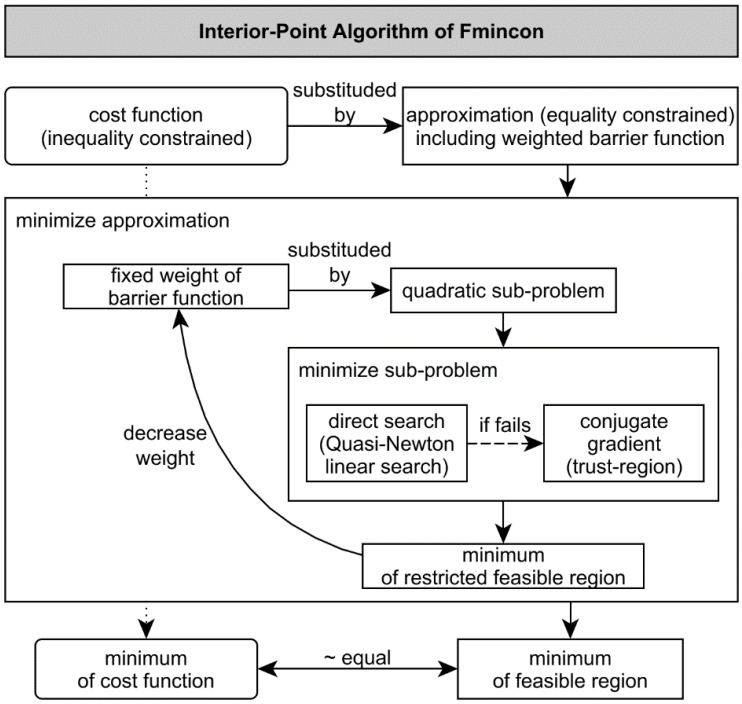


Fig. 3: Scheme of the interior-point method applied by Fmincon (own presentation based on MathWorks (2015b)).

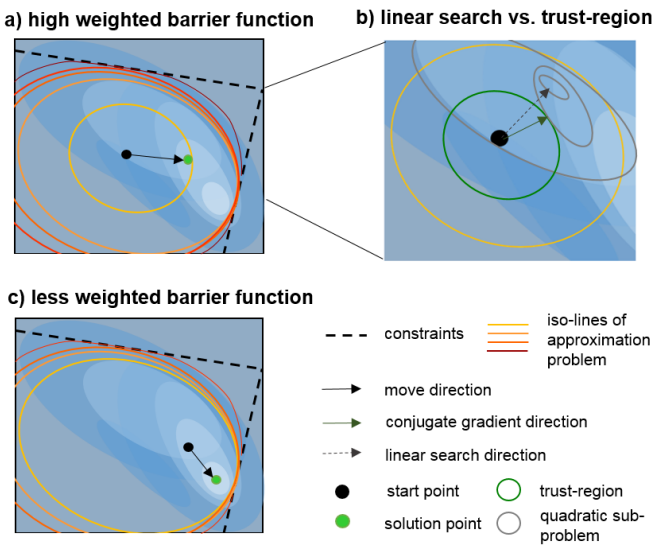


Fig. 4: Illustration of the optimization problems and search directions using Fmincon's interior-point algorithm. Blue background color represents the response surface of the original constraint problem, with bright color indicating the minimum. Black dotted lines show the constraints and limit the region of feasible solutions. Orange-to-red lines illustrate the approximation function and show how the barrier function increases function values near the constraints and thus restricts the feasible region. Starting from a) the algorithm minimizes cost function value within the (restricted) feasible region by either linear search or following the conjugate gradient in a trust region. In b) the difference of the move direction between both approaches is depicted in a schematic manner. While the direct search direction points to the minimum of the quadratic sub-problem (grey lines), the trust-region bounds the area where the quadratic sub-problem is a good approximation of the original response surface. c) shows the next iteration with decreased weight of the barrier function where, starting from the solution point of the previous iteration, the approximation is further minimized (own representation based on MathWorks (2015b)).

After executing the local solver for all feasible start points, the local solutions are ordered by their cost function value and compared with each other with respect to the function and parameter tolerance. Finally, a sorted list of local solutions that have passed the tolerance criteria is reported (MathWorks, 2015a).

Heuristic Approaches

Out of the heuristic methods, popular optimization techniques that generate global solutions are the Genetic Algorithm (GA) (Wang, 1991), the Shuffled Complex Evolution (SCE) algorithm (Duan et al., 1992), Simulated Annealing (SA) (Kirkpatrick et al., 1983), and the Covariance Matrix Adaption Evolution Strategy (CMAES) (Hansen and Kern, 2004). Arsenault et al. (2013) compared different stochastic optimization methods including those three and found SCE performing good for less complex hydrologic models, whereas CMAES outperformed the others in terms of finding the lowest minima and convergence speed for both, small and large parameter spaces.

CMAES is a state of the art evolutionary algorithm to solve non-linear optimization problems. Similar to Quasi-Newton methods, it uses a positive definite matrix within an iterative procedure to estimate the area of interest in which to search for the minimum (Hansen, 2014). In contrast to classical methods CMAES does not rely on derivatives. Instead, it is based on statistical information gathered by repeated sampling of the parameter space. Following the principles of evolution, at each iteration, possible solutions (resp. parameter combinations; individuals) are generated by variation of the previous set of trial solutions (parent population) and selected based on their cost function value (fitness). As shown in Fig. 5, starting from an initial point, a multivariate normal distribution, determined by its mean value and covariance matrix, is used to sample the initial population. Then, each individual of the sampling is ranked by its cost function value and weighted respectively. The weighted combination of the best samples is used to calculate a new mean of the distribution and a new empirical covariance matrix. In general, updating the mean of the distribution maximizes the likelihood of the previous successful individuals, and weighted approximation and adaption of the covariance matrix defines the area of potential low cost function values. The covariance matrix is adapted with the aid of two evolution paths and a step size coefficient. The evolution paths track the progress of the distribution mean (direction and distance) and contain information about the correlation between consecutive steps. One of them, the search path, is used to control the step size, which is reduced if the path is short (small change of the mean) and increased in case of a long path (big change of the mean, indicating best solutions are found in a large distance to the previous mean). The step size then is used to increase or decrease the covariance matrix and thus defines whether the elements of the next sampling are more remote or more close to each

other. The second evolution path that is used to update the covariance matrix includes information about the move direction of the mean. Sampling, evaluation of the fitness and updating the distribution parameter are repeated until a stopping criteria is reached. CMAES then returns the best individual of the last iteration (Hansen and Kern, 2004; Krause, 2007).

As CMAES does not require derivatives and provides feasible results on non-smooth, non-continuous, multi-modal and noisy problems, it performs superior to gradient-based methods if the response surface is rugged, e.g. due to outliers, various local optima or discontinuities. However, because of the adaption process, second-order derivative methods are usually faster, especially for less dimensional problems. Therefore, CMAES is typically applied to unconstrained or bounded constraint problems with 3 to 100 dimensions and when gradient-based methods fail (Hansen and Kern, 2004).

By default, the population size is relatively small, to allow for fast convergence. However, an increase of the population size improves the strategies capability of handling noise and local minima, making CMAES a reliable tool even for global optimization (Hansen, 2014).

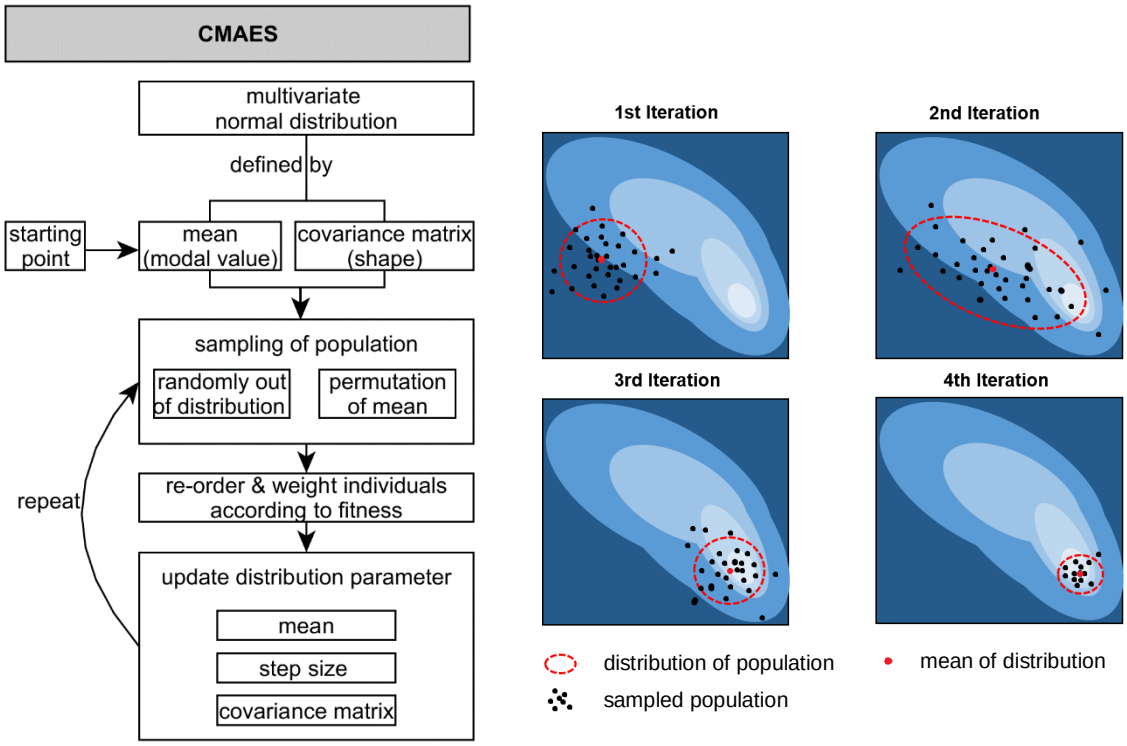


Fig. 5: Scheme of the CMAES algorithm.Left: processing steps; Right: Evolution of distribution and sampled population with each iteration towards the minimum. The different shapes of the distribution of the population illustrate the adaption of the covariance matrix with each step. (population: set of trial solutions, individuals: trial solutions, fitness: cost function value, distribution of population as the iso-density ellipsoid associated with the covariance matrix) (own representation based on Hansen (2014)).

Multi-objective Calibration

Automatic optimization of hydrologic models usually considers one single efficiency criterion, which in principle results in good parametrizations with high reliability (Sorooshian et al., 1993). However, no single-objective function is an adequate measure of all characteristics of the observed data (Yapo et al., 1998). Therefore, multi-objective calibration, which aims to optimize more than one performance criteria, is often more adequate to incorporate all important hydrologic system characteristics (Xu, 2002). Such methods as the Multi-Objective Complex Evolution (Yapo et al., 1998), the Multi-Objective Shuffled Complex Evolution Metropolis Method (Vrugt et al., 2003b) and the Non-dominating Sorting Genetic Algorithm II (Deb et al., 2002) additionally yield the advantage to mitigate effects of parameter equifinality as the number of optimal parameter sets for multiple criteria is less than for solely one criterion (Fischer, 2013). However, the throwback of these approaches is that they do not provide one single solution, but a so called Pareto-front of (equally) good solutions, meaning a set of parameter combinations of which each performs best regarding one criterion. Thus, the selection of the final ‘optimal’ parameter set is based on a subjective compromise between the solutions of the Pareto-front (Fischer, 2013). Additional, computational costs in multi-objective calibration increase with the number of considered efficiency criteria, and in general more iterations are needed than for single-objective optimization (Coello et al., 2007). One way to keep computational costs low, yet to consider different aspects of the hydrological system, is to combine multiple criteria in one cost function instead of minimizing a cost function for each criterion. For instance, Zhang et al. (2008) uses four different objective functions, that cover different aspects of the hydrograph, but optimizes the average of these four.

3 Methods and Data

The following chapter describes the methodologies and data used in this thesis.
Except for individual maps that are created with ArcGIS 10.2.2 (ESRI 2014), the MATrix LABoratory software (MATLAB, TheMathWorks, Inc. 2015) is used for all working steps, calculations and analyses.
An overview on the workflow and how the research questions are addressed is provided by Fig. 6. To find out which model formulation for infiltration and runoff generation is most appropriate and consistent with earth observation data, seven model variants are developed based on hydrological process knowledge and literature review. All model variants are implemented in MATLAB. Simultaneously, global data sets and time series for model forcing and calibration are adjusted to the intended modelling domain and used to delineate the study area. In order to find the most suitable optimization method, four commonly applied search algorithms and various cost functions are tested. Out of them, one method is selected for calibration of the model variants. Thereby calibration intends to find the parameter set that provides optimal agreement between observation and simulations for all grid cells in the study area, meaning a global uniform parameter set. As this approach neglects spatial heterogeneity of physio-geographic conditions, spatialization of runoff-related parameters is incorporated in one model variant by implementing global maps of base flow related characteristics. All calibrated model variants (original seven variants as well as variants with spatialized information) are then evaluated against observed data and compared to each other to identify the most convenient. In the last step, this 'best' model variant is applied and the model output is examined regarding its conformity with terrestrial water storage variations obtained by the GRACE measurements, as well as regarding the relative contributions of snow pack, soil moisture and groundwater to these variations.

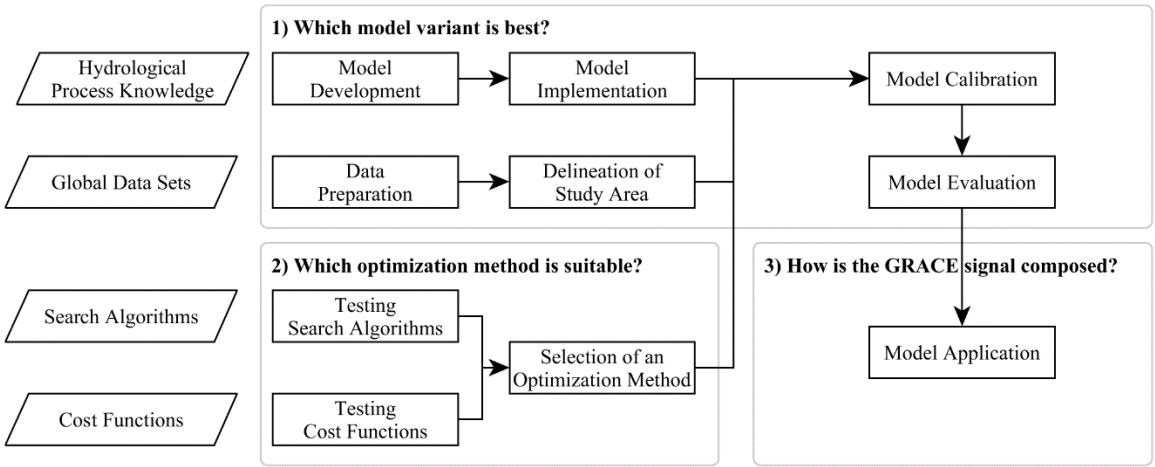


Fig. 6: Workflow and corresponding research questions.

The different model variants and their implementation are explained in detail in the first section of the following chapter. Subsequently, the second section describes the used data sets, their preparation and the delineation of the study area. Finally, the experimental design including the methods used for selection of the most appropriate optimization method, for model calibration and evaluation, as well as for analysis of the model output is illustrated in the last section.

3.1 Model Description

In this study, different model variants that all calculate daily water flows and storages are developed, tested and compared.

All variants are forced by observation based data streams of precipitation (P), air temperature (T), net radiation (Rn) and evapotranspiration (ET). Each model includes descriptions of snow and soil water storages. They simulate the partitioning of P between liquid precipitation (rainfall) and snowfall (SF) as well as snow accumulation, snow melt (M), sublimation from snow cover (ETSub) and the soil water budget. To represent the latter, the model variants use different approaches for allocating incoming water to infiltration (In) and runoff from soil (Qs, surplus water that is not absorbed by soil), depending on the current soil moisture conditions (SM). ET from soil water storage is acquired from input data and, if necessary, adjusted according to available soil moisture.

Further, two methods for the simulation of runoff are implemented. In the basic approach, total runoff (Q) is composed of one component, represented by Qs with consideration of retardation. Introducing this temporal delay allows mimicking the role of storage components other than soil. In the other approach, the models are extended by a separate groundwater component, including groundwater recharge (Perc), groundwater storage (GW) and base flow (Qbase). Thus, they distinguish explicitly between a fast, direct runoff component (Qdir) and delayed Qbase released from groundwater storage. The sum of both represents Q.

Altogether, four approaches to represent the soil water budget and two approaches to simulate Q are implemented. The combination of them results in seven model variants.

Fig. 7 illustrates the general structure of the models, with the extensions of the groundwater variants depicted in gray color. The individual processes and equations are described in detail in the following sections. They are conducted for every model time step and every land surface grid cell. Tab. 2 lists the model variables, which are denoted using Latin capital letters. Parameters (Tab. 3) are written in lower case with a prefix that indicates the corresponding flux, pool or state variable (*p* for scaling parameters, *sn* for parameters affecting snow, *m* for melt process, *s* for soil and *g* for groundwater related processes) and an extension referring to the corresponding modeling approach.

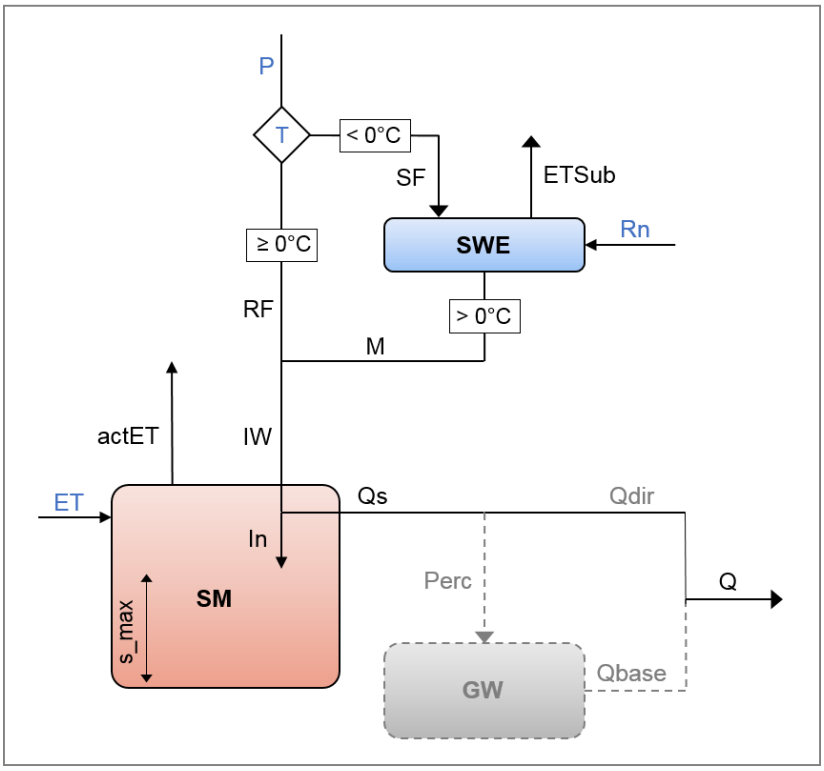


Fig. 7: Schematic structure of the models. Blue color identifies forcing data; gray color identifies extensions of the groundwater model variants.

Tab. 2: Model variables.

Symbol	Description	Unit
Forcing		
P	precipitation	mm d^{-1}
T	temperature	°C
Rn	net radiation	MJ m^{-2} d^{-1}
ET	evapotranspiration	mm d^{-1}
Constants		
G	ground heat flux	MJ m^{-2} d^{-1}
Pair	air pressure	kPa
Pools and State Variables		
FSC	fractional snow cover	-
SWE	snow water equivalent	mm
SM	soil moisture	mm
GW	ground water storage	mm
W	total amount of available water	mm
Fluxes		
SF	snowfall	mm d^{-1}
RF	rainfall	mm d^{-1}
ETSub	sublimation from snow cover	mm d^{-1}
M	melt from snow cover	mm d^{-1}
IW	input water	mm d^{-1}
In	infiltration	mm d^{-1}
Qs	runoff from soil	mm d^{-1}
actET	scaled evapotranspiration	mm d^{-1}
Perc	percolation	mm d^{-1}
Qbase	base flow	mm d^{-1}
Qdir	direct runoff	mm d^{-1}
Q	total runoff	mm d^{-1}

Tab. 3: Model parameters, their description, upper and lower bounds as well as default values.

Parameter	Description	Unit	Min	Max	Default	
Scaling of Meteorological Forcing						
p_sf	scaling parameter for SF	-	0	∞	1	
p_et	scaling parameter for ET	-	0	∞	1	
Snow Component						
m_t	snow melt factor for T	mm K^{-1}	0	10	3	Stacke et al. (2011)
m_r	snow melt factor for Rn	mm MJ m^{-2}	0	3	2	
sn_α	sublimation resistance		0	1	0.95	
sn_c	minimum SWE that ensures complete snow coverage of the grid cell	mm	0	∞	15	Balsamo et al. (2009)
Soil Component						
s_max	maximum soil water holding capacity	mm	0	∞	300	
s_fac_simple	tendency of soil to release water	-	0	1	0.5	Rees et al. (2004)
s_exp_simple	form of relation between runoff and water storage	-	0	∞	1	
s_exp_berg	runoff coefficient	-	0.1	5	1.1	Müller-Schmied et al. (2014)
s_exp_budyko	retention efficiency of soil	-	0	1	0.6	Zhang et al. (2008)
Runoff Component						
Basic variants						
q_t	recession time scale for Qs	d	0	100	2	Orth et al. (2013)
Groundwater variants						
g_r	groundwater recharge rate	-	0	1	0.16	Alley (1984)
g_d	groundwater depletion factor	d^{-1}	0	1	0.01	Müller-Schmied et al. (2014)

3.1.1 Snow Component

Snow storage is implemented as a simple accumulation and melt approach, which further is extended by consideration of sublimation and fractional snow cover. The snow storage as described by the snow water equivalent SWE [mm] at time t [d] is calculated as mass balance:

$$SWE_t = SWE_{t-1} + SF_t - ETSub_t - M_t \tag{13}$$

where SWE_{t-1} [mm] is the snow water equivalent of the preceding time step which is increased by snowfall SF_t [mm d^{-1}] and reduced by the amount of sublimation $ETSub_t$ [mm d^{-1}] and snow melt M_t [mm d^{-1}].

All precipitation P [mm $d^{-}1$] is assumed to fall as snow at temperatures below 0 °C. Since precipitation estimates, especially during the cold season, are known for biases due to substantial under-catch (Rudolf and Rubel, 2005; Seo et al., 2010), *P* is scaled using the parameter ***p_sf*** to derive SF at time t,

$$SF_t = p\_sf \cdot P_t \qquad | \, T < 0 \text{ °C} \tag{14}$$

In order to incorporate sub-grid variability, the fraction of the grid cell covered by snow is computed following the H-TESSEL approach (Balsamo et al., 2009; ECMWF, 2014):

$$FSC_t = \min\left(\frac{SWE_{t-1}}{sn\_c}, 1\right) \tag{15}$$

with fractional snow cover FSC [-] at time t being linearly dependent from SWE_{t-1} of the preceding time step and ***sn_c*** [mm] being the minimum SWE that ensures complete coverage of the grid cell.

Further, snow melt M and sublimation ETSub are assumed to only emerge from snow covered area by using FSC as scaling factor in the calculation of these fluxes.

Snow melt M occurs when snow storage is present and temperature exceeds melting temperature. Based on the restricted degree-day radiation balance approach described by Kustas et al. (1994), melt M [mm d^{-1}] at time t depends on temperature T_t [°C] and net radiation Rn_t [MJ m^{-2} d^{-1}]:

$$M_t = (m\_t \cdot T_t + m\_r \cdot Rn_t) \cdot FSC_t \qquad | \, T > 0 \text{ °C} \tag{16}$$

where the degree-day factor ***m_t*** [mm °C^{-1}] and the radiation factor ***m_r*** [mm MJ^{-1}] control the melt rate. ***m_t*** usually ranges from 0 to approximately 5 mm °C^{-1} (Julander, 2000), while based on the enthalpy of fusion of ice (0.3337 MJ mm^{-1}), the physical upper limit of ***m_r*** is 3 mm MJ^{-1}.

The derivation of snow sublimation ETSub is adapted from the approach implemented in the GLEAM model. This technique is based on the Priestley and Taylor (1972) formula, which calculates evaporation rate as latent heat flux LE [MJ m^{-2} d^{-1}] based on the available energy Rn [MJ m^{-2} d^{-1}], ground heat flux G [MJ m^{-2} d^{-1}]) and a dimensionless coefficient ***sn_α*** that parameterizes evaporation-resistance. LE at time t is derived by

$$LE_t = \left(sn\_\alpha \cdot \frac{\Delta_t}{\Delta_t + \gamma_t} \cdot (Rn_t - G)\right) \cdot FSC_t \tag{17}$$

with Δ_t being the slope of the temperature/saturated vapor pressure curve [kPa K^{-1}] and γ_t representing the psychrometric constant [kPa K^{-1}]. Both, Δ and γ, are modified for snow covered areas according to Murphy and Koop (2005).
They calculate Δ_t as a function of T_t [K] (Eq. (18)), and γ_t as a function of atmospheric pressure Pair [kPa], specific heat of air at constant pressure c_p [MJ kg^{-1} K^{-1}], the ratio molecular weight of water vapor/dry air MW and latent heat of sublimation of ice λ [MJ kg^{-1}] (Eq. (19)).

$$\Delta_t = \left(\frac{5723.265}{T_t^2} + \frac{3.53069}{T_t - 0{,}00728332}\right) \cdot e^{9.550426 - \frac{5723.265}{T_t} + 3.53068 \cdot \ln(T_t) - 0{,}00728332 \cdot T_t} \tag{18}$$

$$\gamma_t = \frac{Pair \cdot c_p}{MA \cdot \lambda_t} \tag{19}$$

In Eq. (19), Pair is assumed to be time- and space-invariant with a uniform value of 101.3 kPa by what c_p = 0.001 MJ kg^{-1} K^{-1}. MA is a constant of 0.622 and λ is defined by Murphy and Koop (2005) as a function of T_t [K]. With a molecular mass of water of 18.01528 g mol^{-1}, λ can be calculated as:

$$\lambda_t = \left(46782.5 + 35.8925 \cdot T_t - 0.07414 \cdot T_t^2 + 541.5 \cdot e^{-\left(\frac{T_t}{123.75}\right)^2}\right) \cdot \frac{0.001}{18.01528} \tag{20}$$

Since snow-covered ecosystems can be assumed to be unstressed due to the sufficient availability of water, LE corresponds to actual sublimation ETSub (Miralles et al., 2011). And ETSub [mm d^{-1}] can be converted from LE through division by λ:

$$ETSub_t = \frac{LE_t}{\lambda_t} \tag{21}$$

Altogether, the current model calculates ETSub as a function of T_t, Rn_t, Pair, G, ***sn_α*** and FSC_t. While T_t, Rn_t and FSC_t are variable in time and space and depend on input data, the approach postulates constant Pair and G with values of 101.3 kPa resp. 0 MJ m^{-2} d^{-1}.

3.1.2 Soil Component

The central part of the model is the soil water component, which distributes input from rain fall and snow melt to soil water storage SM [mm], actual evapotranspiration actET [mm d^{-1}] and runoff from soil Qs [mm d^{-1}]. The latter can be understood as surplus water, which does not increase soil water storage.

Like snow, the calculation of soil water storage as represented by soil moisture SM [mm] at time t [d] follows the mass balance

$$SM_t = SM_{t-1} + In_t - actET_t \tag{22}$$

with SM_{t-1} [mm] being the soil moisture of the preceding time step which is increased by infiltration In_t [mm d^{-1}] and reduced by actual evapotranspiration $actET_t$ [mm d^{-1}].

On the one hand, the amount of infiltration In is determined by the inflow IW [mm d^{-1}], which is derived from rainfall RF (precipitation P if T ≥ 0°C) and snow melt M at time t [d]:

$$IW_t = RF_t + M_t \tag{23}$$

On the other hand, a part of IW may not infiltrate but contribute to runoff from soil Qs. In this study, four approaches to estimate the partitioning of IW into SM and Qs are tested regarding their suitability for a 1° x 1° daily model.

Except the first, all of them assume a maximum water holding capacity of the soil layer ***s_max*** [mm] and relate it to the available amount of water in the soil to calculate either In or Qs. Respectively, the other flux is derived in accordance to the law of conservation of mass, so that IW equals the sum of Qs and In at time t [d]:

$$IW_t = In_t + Qs_t \tag{24}$$

The explicit formulations used in this study are introduced in the following.

Simple Approach

Based on the strongly simplified assumption that discharge solely depends on the total water storage (Brutsaert and Nieber, 1977; Kirchner, 2009), the first approach considers one single conceptual storage that comprises all storage components as soil moisture and groundwater. Hereinafter it is referred to as ***simple approach*** and its variables and parameters are marked by the extension ***simple***.

Following Brutsaert and Nieber (1977) and Rees et al. (2004) outflow is assumed to be proportional to a power of the storage, which yields the advantage to only need a minimum number of parameters for runoff estimation (Melsen et al., 2014). Accordingly, the method estimates Qs [mm d^{-1}] at time t [d] as a function of the total amount of available water W [mm] and the calibration parameters ***s_fac_simple*** and ***s_exp_simple***:

$$Qs_t = s\_fac\_simple \cdot W_t^{s\_\exp\_simple} \tag{25}$$

s_fac_simple describes the tendency of soil to release water and its range varies between 0 – 1. A larger value implies more water contributing to Qs, with ***s_fac_simple*** = 1 resulting in complete conversion to Qs. The unitless parameter ***s_exp_simple*** ranges from 0 - ∞ and ensures that the function is strictly monotonically increasing, so that Qs increases with W. Further, ***s_exp_simple*** describes the form of storage, e.g. ***s_exp_simple*** = 1 indicates a linear storage and ***s_exp_simple*** = 2 a quadratic one (Rees et al., 2004).

As total amount of available water W includes inflow as well as soil and groundwater storages, W [mm] at time t [d] is the sum of inflow IW and the soil moisture SM of the preceding time step:

$$W_t = SM_{t-1} + IW_t \tag{26}$$

Hence, no explicit separation between soil moisture and groundwater storage is possible and no groundwater model variant can be formulated for the simple approach.

Saturation Approach

The second approach is consistent with the bucket-scheme of Manabe (1969). All inflow is allocated to soil water storage *SM* until the maximum water holding capacity (field capacity)

s_max is reached. Only surplus water contributes to Qs. This method is referred to as ***saturation approach*** and marked by the extension ***sat***.

In [mm d^{-1}] at time t [d] is derived by comparing the possible inflow IW with the available storage capacity ***s_max*** – SM_{t-1} [mm]:

$$In_t = \min(IW_t, s\_\max - SM_{t-1}) \tag{27}$$

If the inflow is less than the available storage capacity, the total amount of inflow is allocated to the soil storage so that In = IW and Qs = 0. If the inflow exceeds the available storage capacity, In = ***s_max*** - SM_{t-1} and the remaining amount of inflow contributes to Qs, with Qs = IW if the soil is completely saturated.

Bergström Approach

The third approach is adopted from the WaterGAP 2.2 model (Alcamo et al., 2003; Döll et al., 2003) that calculates Qs [mm d^{-1}] at time t [d] after Bergström (1995) as:

$$Qs_t = IW_t \cdot \left(\frac{SM_{t-1}}{s\_max}\right)^{s\_\exp\_berg} \tag{28}$$

In Eq. (28) Qs depends on the inflow IW, the runoff coefficient ***s_exp_berg*** and the actual soil moisture SM compared to its maximum water holding capacity ***s_max***. As in the saturation approach, Qs = 0 if the soil water storage is empty and Qs = IW if the soil is completely saturated. However, between these points, ***s_exp_berg*** determines the amount of inflow that converts to Qs as illustrated in Fig. 8a. While low values of ***s_exp_berg*** lead to a high amount of Qs even if the soil moisture deficit is low (e.g. low SM/***s_max*** ratio***)***, higher values of ***s_exp_berg*** increase the proportion of IW that infiltrates.

This approach is henceforth referred to as ***Bergström approach*** and denoted using the extension ***berg***.

Budyko Approach

Finally, the forth approach is based on a generalized demand and supply framework as postulated by Zhang et al. (2008). Their attempt relies on the method of Budyko (1961) for modeling mean annual evapotranspiration using a demand/supply index and a similar concept by Fu (1981). Both were adapted to simulate the partition of precipitation into direct runoff and rainfall retention. In the context of this study, direct runoff corresponds to runoff from soil Qs and rainfall retention to infiltration In. The demand limit for In_t is the available storage capacity ***s_max*** – SM_{t-1}, while the supply limit can be considered as inflow IW_t.

The ratio (***s_max*** – SM_{t-1})/IW_t then represents the demand/supply index for the partition of IW. In [mm d^{-1}] at time t [d] is calculated using Fu's equation:

$$In_t = IW_t \cdot \left\{ 1 + \frac{\boldsymbol{s\_max} - SM_{t-1}}{IW_t} - \left[1 + \left(\frac{\boldsymbol{s\_max} - SM_{t-1}}{IW_t} \right)^{\frac{1}{1-\boldsymbol{s\_exp\_budyko}}} \right]^{1-\boldsymbol{s\_exp\_budyko}} \right\} \qquad (29)$$

where ***s_exp_budyko*** represents the retention efficiency. The parameter ranges from 0 – 1, with a larger value resulting in less Qs and more In as illustrated in Fig. 8b. Fig. 8b further shows, that considering low parameter values, small variations lead to considerable differences in the amount of simulated Qs resp. In. In comparison, for ***s_exp_budyko*** > ~0.4 the differences are less pronounced and the relation between Qs/IW and SM/***s_max*** appears almost linear if a certain soil moisture is reached.

In the following, the fourth approach is referred to as ***Budyko approach*** and is marked by the extension ***budyko***.

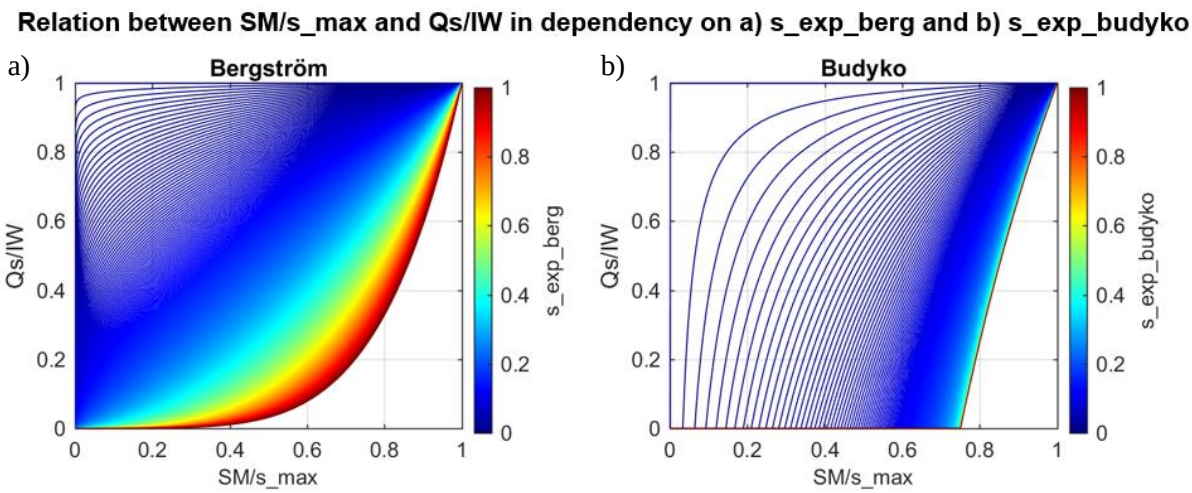


Fig. 8: Influence of the parameter value of a) s_exp_berg and b) s_exp_budyko on the proportion of IW that contributes to Qs depending on soil moisture conditions.(In b) s_max = 300 mm is assumed).

Evapotranspiration

Since in this study an evapotranspiration product is utilized instead of calculating its components, the computed In is added to soil water storage before accounting for ET. Thus, Eq. (22) is split during the modeling procedure into Eq. (30) and Eq. (33).

First, soil moisture SM [mm] of the preceding time step t-1 [d] is increased by infiltration In [mm d^{-1}]:

$$SM_{temp} = SM_{t-1} + In_t \qquad (30)$$

When considering evapotranspiration, the input data is scaled similarly to precipitation to provide for data inconsistency. Thus, actual evapotranspiration actET [mm d^{-1}] at time t [d] is the product of evapotranspiration ET [mm d^{-1}] and the scaling parameter ***p_et***:

$$actET_t = p\_et \cdot ET_t \tag{31}$$

To ensure that actET does not exceed the amount of water available in the soil, it is compared to the actual soil moisture SM_{temp} (Eq. (32)) before decreasing SM_{temp} and determine the final SM at time t (Eq. (33)).

$$actET_t = \min(actET_t, SM_{temp}) \tag{32}$$

$$SM_t = SM_{temp} - actET_t \tag{33}$$

3.1.3 Runoff Component

For modelling how the surplus water Qs is transferred to actual runoff Q, two approaches are investigated, namely the ***basic*** and the ***groundwater*** variant.

Basic Runoff Approach

In the basic approach, the whole amount of runoff from soil contributes to discharge. As this one runoff component comprises fast direct runoff as well as delayed interflow and base flow, Orth et al. (2013) found consideration of retardation useful. Accordingly, total discharge Q [mm d^{-1}] at time t [d] results from the accumulated effects of all runoff from soil Qs [mm d^{-1}] generated during the preceding 60 time steps:

$$Q_t = \sum_{i=0}^{60} Qs_{t-i} \cdot \underbrace{\left[e^{-\frac{i}{\mathbf{q\_t}}} - e^{-\frac{i+1}{\mathbf{q\_t}}}\right]}_{\text{delay component}} \tag{34}$$

where the recession time scale ***q_t*** [d] determines how quickly runoff from soil is transformed into streamflow. In theory, an infinite number of time steps would be necessary to ensure that all generated Qs is transformed into discharge. However, the arbitrary number of 60 days allows accounting for > 99 % of Qs (Orth et al., 2013), as long as ***q_t*** is below 13 days. To allow longer recession times when calibrating the model and still account for > 99 % of Qs within the 60 days-window, the delay component of Eq. (34) is scaled with its sum.

Introducing temporal delay leads to retention of a portion of Qs. Assuming that this portion is retained solely in groundwater, the change of groundwater storage ΔGW [mm d^{-1}] at time t [d] can be inferred indirectly using the water balance:

$$0 = P_t - actET_t - Q_t + \Delta TWS_t \tag{35}$$

with the change of total water storage ΔTWS [mm d^{-1}] resulting from

$$\Delta TWS = (SWE_t - SWE_{t-1}) + (SM_t - SM_{t-1}) + \Delta GW_t \tag{36}$$

so that solving Eq. (35) and Eq. (36)

$$\Delta GW_t = actET_t + Q_t - P_t - (SWE_t - SWE_{t-1}) - (SM_t - SM_{t-1}) \tag{37}$$

Groundwater Approach

In the second runoff approach tested, a separate linear storage layer is implemented to account for groundwater and delayed base flow explicitly. This attempt is similar to common GHMs and the simple conceptual abcd-model as introduced by Thomas Jr (1981).

Analogous to snow and soil, the groundwater storage GW [mm] at time t (d) is calculated by mass balance as:

$$GW_t = GW_{t-1} + Perc_t - Qbase_t \tag{38}$$

where the previous groundwater storage GW_{t-1} [mm] is increased by percolation $Perc_t$ [mm d^{-1}] and decreased by base flow $Qbase_t$ [mm d^{-1}].

Percolation Perc [mm d^{-1}] at time t [d] is calculated as a fraction ***g_r*** of runoff from soil Qs_t [mm d^{-1}]:

$$Perc_t = g\_r \cdot Qs_t \tag{39}$$

where the parameter ***g_r*** represents the groundwater recharge rate which can vary between 0 (no groundwater recharge) and 1 (all runoff from soil is allocated to the groundwater).

Base flow Qbase [mm d^{-1}] at time t [d] depends on groundwater storage GW_t [mm]:

$$Qbase_t = g\_d \cdot GW_t \tag{40}$$

where the groundwater depletion factor ***g_d*** [d^{-1}] corresponds to the reciprocal of the groundwater residence time. Similar to ***g_r***, the range of ***g_d*** is 0 – 1.

To identify GW [mm] at time t [d] the mass balance (Eq. (38)) can be solved using Eq. (40):

$$GW_t = \frac{1}{1 + g\_d} \cdot (GW_{t-1} + Perc_t) \tag{41}$$

Fast direct discharge Qdir [mm d^{-1}] at time t [d] equals the amount of Qs [mm d^{-1}], which does not contribute to groundwater recharge so that:

$$Qdir_t = (1 - \boldsymbol{g\_r}) \cdot Qs_t \tag{42}$$

The sum of direct discharge Qdir [mm d^{-1}] and base flow Qbase [mm d^{-1}] represents total discharge Q [mm d^{-1}] at time t [d]:

$$Q_t = Qdir_t + Qbase_t \tag{43}$$

3.1.4 Model Implementation

The above described approaches are implemented in MATLAB 2015a. Each model component is defined as a separate function, of which the final model variants are assembled as shown in Fig. 9. As illustrated, the snow component is common for all model variants and is characterized by five parameters. The soil component comprises two to three parameters depending on the approach, and the runoff component applies one (basic version) resp. two (groundwater version) parameters. Thus, the total number of model parameters ranges from eight (SatBasic) to ten (BergGW, BudykoGW).

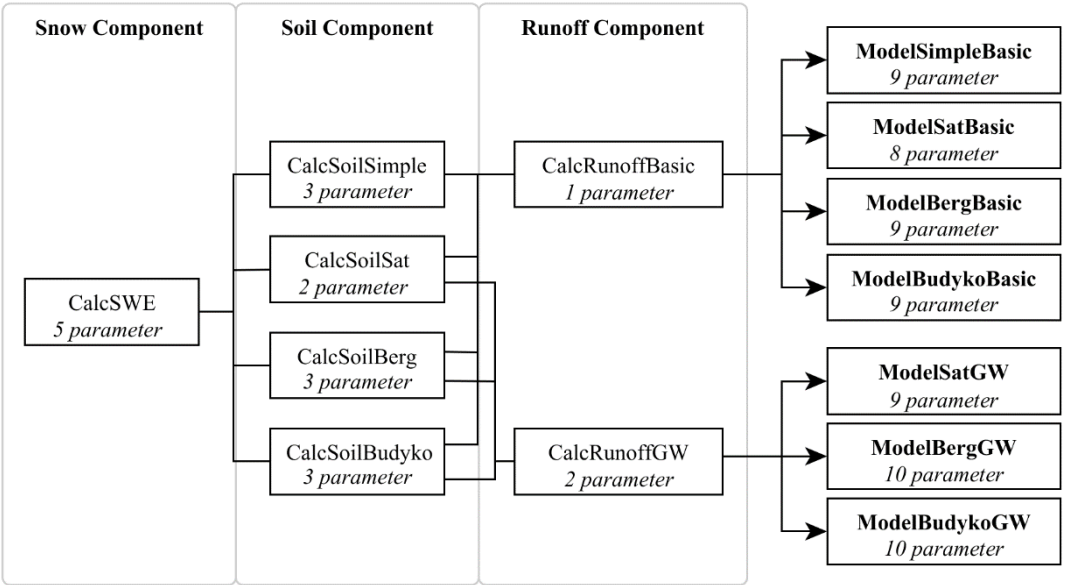


Fig. 9: Model configuration in MATLAB(Calc indicates a MATLAB function, while Model refers to the final combination of these functions) and number of parameters.

3.2 Data

This study uses daily time-series of globally gridded precipitation, temperature, net radiation and evapotranspiration data products as meteorological forcing. Further, data acquired in the Gravity Recovery and Climate Experiment (GRACE) and in the GlobSnow Project are utilized for model optimization and evaluation. Temporally constant spatial data is used to define the study area and to delineate climate zones for spatial distributed evaluation. Additionally, regional information related to base flow is tested for model refinement.

An overview on the data products and their main characteristics is given in Tab. 4.

Tab. 4: Overview on input data (NH: Northern Hemisphere).

Data	Data Basis	Coverage & Resolution		Reference
		Spatial	Temporal	
Meteorological Forcing				
Precipitation	GPCP v1.2	1° x 1° global	daily 1996 - present	Huffman et al. (2001), Huffman and Bolvin (2013)
Temperature	WFDEI	0.5° x 0.5° global	daily 1979 - 2012	Weedon et al. (2014)
Net Radiation	CERES	1° x 1° global	3-hourly 03/2000-05/2015	Wielicki et al. (1996)
Evapotranspiration	FLUXNET	0.5° x 0.5° global	daily 1982 - 2013	Tramontana et al. (2016)
Optimization and Evaluation				
Terrestrial Water Storage Anomalies	GRACE Tellus JPL- RL05M	0.5° x 0.5° global	monthly 2002 - 2015	Wiese et al. (2015)
Snow Water Equivalent	GlobSnow v2.0	0.25° x 0.25° non-alpine NH	daily 1979 - 2012	Luojous et al. (2014)

Continuation of Tab. 4.

Ancillary Data				
Water mask	MODIS land cover MCD12Q1	0.0083° x 0.0083° global	-	
regional base flow information (BFI1-4, K)	Global Maps of Streamflow characteristics (related to base flow) v1.9	0.125° x 0.125° global	-	Beck et al. (2015)
Köppen-Geiger Climate Zones	World Map of Köppen-Geiger Climate Classification	0.5° x 0.5° global	-	Kottek et al. (2006)

3.2.1 Meteorological Forcing Data

Precipitation

Daily precipitation sums represent the main model drivers. In this study, the 1° daily precipitation product version 1.2 of the Global Precipitation Climatology Project (GPCP-1DD) is used (Huffman et al., 2001; Huffman and Bolvin, 2013). This dataset provides daily, globally gridded values of precipitation totals since October 1996, which are produced by a combination of precipitation estimates from remote sensing and precipitation gauge analysis.

Between 40° N to 40° S, precipitation is estimated by the Threshold-Match-Precipitation Index (TMPI). This approach uses infrared brightness temperatures which are measured by a variety of sensors aboard geosynchronous satellites as the Geosynchronous Operational Environmental Satellites (GEOS, United States) and the Meteorological Satellite (Meteosat, European Community). The 3-hourly measured temperatures are compared with a threshold temperature and a conditional rain rate that are set locally by month. For definition of these parameters, Special Sensor Microwave Imager-based (SSMI) resp. Special Sensor Microwave Imager Sounder (SSMIS, since late 2003) precipitation frequencies and the GPCP satellite-gauge combined monthly precipitation estimate (GPCP-SG) are used (Huffman et al., 2001). The latter, as the daily product, is derived from a precipitation estimate merged from the same infrared, microwave and sounder satellite data, which then is linearly combined with gauge analysis (Huffman and Bolvin, 2013). The in situ component of this analysis is the Monitoring Product provided by the Global Precipitation Climatology Centre (GPCC) that is based on SYNOP and monthly CLIMAT reports from 7000 - 8000 stations (Schneider et al., 2015).

As the viewing geometry of the geosynchronous infrared sensors becomes unfavorable near the limb of the satellite's view, sounding data from low-earth polar orbit satellites is used to estimate precipitation at latitudes higher than 40°. In this area, the Television and Infrared Observation Satellite Operational Vertical Sounder (TOVS, until 2005) and the Atmospheric Infrared Sounder (AIRS, since 2005) precipitation datasets are used. They are rescaled to match

the TMPI precipitation estimates at the data boundaries (39-40°). The resulting non-zero values are further scaled locally to sum to the monthly GPCP-SG (Huffman et al., 2001).
According to Crow (2007), GPCP-1DD represents one of the best available global precipitation products and has been widely used in different studies. As a gauge-corrected product, it is likely to outperform fully satellite-based products, especially in areas with dense observational-networks (Miralles et al., 2011).

Temperature

Mean daily temperature is obtained from the WFDEI data product (Weedon et al., 2014). WFDEI is generated by applying the Water and Global Change (WATCH) forcing data methodology to the ERA-Interim reanalysis data (Dee et al., 2011). The dataset has been created as part of the European Union (EU) WATCH project (Harding et al., 2011) and provides global meteorological forcing data for land area on a 0.5° x 0.5° grid for the period 1979 – 2012 (Weedon et al., 2014).
To generate the WFDEI temperature product, the ERA-Interim temperature data was adjusted such that their monthly means match the monthly temperature dataset from CRU (Rust et al., 2015). The CRU time-series themselves are based on observations from more than 4000 weather stations worldwide that are interpolated to a 0.5° x 0.5° global grid (Jones and Harris, 2008). Further, the adjusted data is elevation corrected by applying an environmental lapse rate. The resulting daily averaged WFDEI product represents air temperature in Kelvin at 2 m above the surface.
As WFDEI is aimed to support large-scale hydrological modeling, it's most likely appropriate for this study.
In order to use the WFDEI temperature data as input for a 1° x 1° gridded model, a resampled version of the dataset was provided by the MPI-BGC.

Net Radiation

Net radiation is defined as the balance between incoming and outgoing shortwave and longwave radiation fluxes at the top of the earth's atmosphere (Sai Krishna et al., 2014). It describes the total amount of energy that is available to influence the climate, and thus, among others, processes as evaporation (Pan et al., 2015).
In this study, net radiation is estimated from the SYN1deg Ed3A data product of the Clouds and the Earth's Radiant Energy Systems (CERES) program of the United States' National Aeronautics and Space Administration (NASA) (Wielicki et al., 1996). This dataset represents one of the few sources of currently available Earth Radiant Budget data (Sai Krishna et al., 2014). It provides 3-hourly and monthly shortwave and longwave radiation fluxes from March 2000 to May 2015 with 1° grid spacing.

The product is generated from top-of-atmosphere radiance observed by the CERES instrument onboard the NASA's Terra (since 03/2000) and Aqua (additional since 07/2002) satellites, combined with coincident cloud and aerosol properties derived from the Moderate-resolution Imaging Spectroradiometer (MODIS), as well as cloud properties and broadband fluxes obtained from different geostationary satellites (NASA, 2013).

Using the daily averaged CERES SYN1deg Ed3A data, net radiation (Rn) is calculated with the principle of energy balance from its components, including downward and upward shortwave radiation (RS↓ resp. RS↑) as well as downward and upward longwave radiation (RL↓ resp. RL↑) (Sai Krishna et al., 2014) :

$$Rn = (RS\downarrow - RS\uparrow) + (RL\downarrow - RL\uparrow) \tag{44}$$

Evapotranspiration

Evapotranspiration forcing is based on estimates of latent energy *LE* derived by using a data-driven approach as described in Tramontana et al. (2016). The procedure integrates eddy covariance measurements of land-atmosphere exchanges at globally distributed sites of the FLUXNET and CarboAfrica network (Valentini 2014), remote sensing data from MODIS as well as meteorological data in a machine learning algorithm. The resulting dataset provides daily estimates of LE in Mega Joule per square meter and day on a 0.5° x 0.5° global grid for the period 1980 – 2013.

To upscale in-situ observations to global scale, the approach utilizes a random forest algorithm that is trained on site level using

a) daily meteorological data from the sites combined with long-term time series of CRU-NCEP v6 (Vivoy, 2015), which is a merged product of CRU monthly 0.5° climate variables (1901-2013) (New et al., 2000) and a 6-hourly 2.5° National Centers for Environmental Prediction (NCEP) reanalysis dataset (1948-2013) (Kalnay et al., 1996),
b) the mean seasonal cycle (MSC) of MODIS data averaged for the 3x3 km surrounding area of each site and interpolated to daily time steps,
c) additional variables derived by combining a) and b), including MSC and corresponding metrics of dynamic variables as well as plant functional type (PFT) derived from the majority class of the MODIS land cover product and a water availability index (WAI) based on a simple dynamic water balance model (Tramontana et al., 2016).

Out of the pool of potential predictor variables, the most suitable subset is chosen by applying a Guided Hybrid Genetic Algorithm (GHGA) (Jung and Zscheischler, 2013).

The random forest is trained using 10-fold cross-validation and subsequently forced with the global data sets of predictor variables tiled by PFT. Therefor the algorithm is run separately for

grids of each PFT consisting of the average value per PFT and time step at 0.5° and finally the weighted mean over the PFT fractions is calculated for each grid cell (Tramontana et al., 2016).

3.2.2 Calibration and Validation Data

GRACE Data

Model optimization in this study mainly relies on terrestrial water storage anomalies obtained from the Gravity Recovery and Climate Experiment (GRACE). GRACE as a joint satellite mission between the NASA and the German Aerospace Center (DLR) provides time-varying measurements of the Earth's gravitational field since April 2002 by making accurate measurements of the inter-satellite range between two co-planar, low-altitude near-polar orbiting satellites (GRACE-A and GRACE-B) using GPS and a microwave ranging system (Kruizinga and Williams, 2010). Variations in the Earth's gravitational field represent changes in mass distribution, and thus, after removal of processes due to atmospheric and ocean dynamics, solid earth and ocean tides, primarily changes in terrestrial water storage.

In this study, the GRACE Tellus Mascon product based on the GRACE gravity fields Release 05 processed at NASA's Jet Propulsion Laboratory (JPL) (Watkins and Yuan, 2012) is applied (Wiese et al., 2015). The dataset provides anomalies of equivalent water thickness in centimeter relative to the January 2004 – December 2009 time-mean baseline for both land and ocean. The monthly time-series, as well as local scaling coefficients for land area and uncertainty estimates in centimeter, are available on a 0.5° x 0.5° grid (Watkins et al., 2015; Wiese et al., 2015).

Unlike previous GRACE Tellus products, the JPL RL05M dataset is derived by using equal-area 3x3° spherical cap mass concentration blocks (mascons) to solve for monthly gravity field variations. In comparison to spherical harmonic coefficients, this approach allows for easier implementation of geophysical constraints to filter out noise from the JPL Level-2 Product. Further, a Coastline Resolution Improvement (CRI) filter has been applied to ensure a clean separation along coastlines within land/ocean mascons (Watkins et al., 2015). For each mascon, uncertainties are estimated by scaling the formal covariance matrix, so that, over land, they roughly match uncertainty estimates simulated with methods described by Wahr et al. (1998). Due to the application of a Kalman filter that links adjacent months in the data stream, uncertainty estimates are slightly higher at the beginning and the end of the time series (Watkins et al., 2015).

GRACE products provide a unique homogeneous, global data source that offers insight into regional and continental-scale hydrologic processes and is widely used in current scientific research (Forman et al., 2012). As the monthly estimates of variations in terrestrial water storage include changes of snow, ice, surface water, soil moisture and groundwater, the dataset

is appropriate for calibration and validation of large-scale hydrologic models as developed in this study (Sakumura et al., 2014).

GlobSnow Data

The second dataset used for model optimization in this study is the Snow Water Equivalent (SWE) v2.0 estimate generated as part of the European Space Agency's (ESA) GlobSnow project (Luojous et al., 2014). SWE as the product of snow depth and snow density represents the resulting water column in case a snow pack melts in place, and thus strongly influences the water cycle and biogeochemical cycling in mid- and high-latitudes and high elevation areas (Takala et al., 2011). GlobSnow SWE represents a satellite-based, daily dataset for the period 1979-2012. It provides SWE estimates [mm] for the non-alpine land surface area of the Northern Hemisphere with the exception of Greenland on a 0.25° x 0.25° global grid (Derksen et al., 2014).

The processing scheme applies passive microwave measurements from the Scanning Multichannel Microwave Radiometer (SMMR) and SSMI/S sensors onboard the Nimbus-7 resp. DMSP F8, F11, F13, and F17 satellites, and weather station observations of snow depth collected by the ECMWF in an assimilation scheme described in Takala et al. (2011). The processing utilizes the Helsinki University of Technology (HUT) snow emission model, which describes brightness temperature (T_B) measured by microwave sensors as a function of SWE, snow density and snow grain size (Pulliainen et al., 1999). In a first iteration, the model is used to derive grain size estimates at the locations of weather stations from spaceborne T_B, a constant snow density of 0.24 g cm^{-3} and snow depth observations. Punctual snow depth observations and grain size estimates are interpolated by kriging to form a continuous field and then are used as input for the HUT model to retrieve SWE through forward simulations of T_B. Hereby, the assimilation adaptively weighs the satellite T_B with the observed snow depth map to estimate the final SWE (Takala et al., 2011).

Validation of the weekly aggregated GlobSnow SWE product against independent ground-based measurements from Russia resp. the former Soviet Union revealed an average RMSE around 32 mm for reference SWE values below 150 mm, and an increasing retrieval uncertainty when SWE is above this threshold. Thus, when considering the full reference datasets, the RMSE is approximately 43.5 mm (Luojous et al., 2014). Similar results are obtained for reference data from Finland and Canada and relate to a systematically underestimation of SWE by passive microwave retrieval algorithms under deep snow conditions. This well documented underestimation occurs due to a change in the microwave behavior of the snow pack, which transitions from a scattering medium into a source of emission (Takala et al., 2011). Fig. 10 shows the comparison of GlobSnow SWE estimates and ground measurements of 7388 stations across the Northern Hemisphere by Liu et al. (2014), and illustrates this underestimation.

Additionally, Fig. 10 suggests a tendency of the GlobSnow product to overestimate SWE values less than 30 mm.

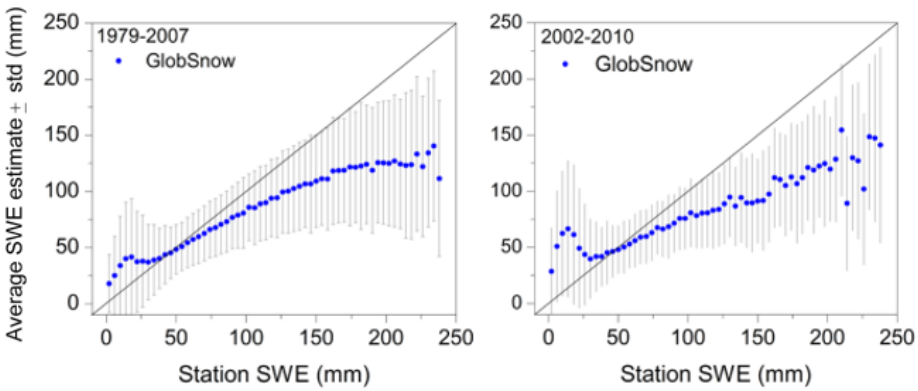


Fig. 10: Comparison of monthly mean GlobSnow SWE estimate and ground-measured SWE from 7388 meteorological stations across the northern hemisphere (Global Historical Climatology Network, NOAA) with a) 139 877 samples and b) 21 501 samples (Liu et al., 2014).

However, GlobSnow provides a clear improvement in retrieval accuracy compared to pure passive microwave earth observation products that already saturate for reference SWE values between 80 – 120 mm (Hancock et al., 2013; Liu et al., 2014). Thus, the GlobSnow SWE as an assimilation product exploits the benefits of both conventional and passive microwave data, and shows superior performance in comparison to stand-alone approaches (Derksen et al., 2014). Additionally, its hemisphere coverage makes it the most suitable data set currently available for testing snow processes within large-scale hydrological models (Hancock et al., 2013).

3.2.3 Ancillary Data

Water Mask

For delineation of the study area and reducing issues due to land/water mixed pixel during the modeling, a water mask is applied. The mask provided by Ichii Kazuhito (Fukushima University) is derived from the global MODIS land cover product MCD12Q1 based on data from the year 2001 (Friedl et al., 2010). The dataset has been resampled from originally 500 km spatial resolution to 1 km (0,0083°) by taking the majority class. A further resampling to 1° x 1° has been conducted at the MPI-BGC. The mask contains information on the fraction of water area within one grid cell, ranging from 0 (no water surface area) to 1 (surface completely covered by water).

Base flow related Streamflow Characteristics

Spatial distributed information related to base flow is obtained from global maps of streamflow characteristics produced by Beck et al. (2015). They trained neural network ensembles with

observed streamflow from 3000 – 4000 catchments (10 – 10000 km^2) to estimate a total number of 17 characteristics based on climate and physiographic features. The final maps are derived by applying the trained networks globally on ice-free land. They have a spatial resolution of 0.125° x 0.125° and represent the median of the neural network ensembles for each characteristic.

In this study, the base flow index computed by using four different techniques (BFI1-4) and the base flow recession constant (K) come into use. The latter represents the rate of base flow decay, whereas BFI is defined as the ratio of the long-term base flow to total streamflow. They are calculated from daily streamflow by applying different methodologies as summarized in Tab. 5. It should be noted, that Beck et al. (2015) define base flow "as the slowly varying portion of [streamflow that] includes all slow runoff components".

Tab. 5: Streamflow characteristics related to base flow as computed and provided by Beck et al. (2015).

Characteristic	Unit	Computation
BFI1	-	recursive digital filter, window size of 5 days (Van Dijk, 2010)
BFI2	-	local-min method, duration of surface-runoff set to 5 (Pettyjohn and Henning, 1979; Sloto and Crouse, 1996)
BFI3	-	7-day sliding interval (7-day moving min) (Pettyjohn and Henning, 1979; Sloto and Crouse, 1996)
BFI4	-	subsequently connecting valleys in series of minima at 5-day non-overlapping intervals (Gustard et al., 1992)
K	d^{-1}	window size of 5 days, ignoring days with zero flow (Beck et al., 2013)

Köppen-Geiger Climate Zones

For spatial distributed evaluation of the models' performance, the updated digital world map of Köppen-Geiger climate classification presented by Kottek et al. (2006) is used to partition the study area. The map is based on the effective climate classification by Köppen (1900), who distinguishes five vegetation groups and applies hygral and thermal thresholds for further subdivision. Accordingly, the five main climates (A – equatorial zone, B – arid zone, C – warm temperate zone, D – snow zone, E – polar zone) are specified by a second letter referring to precipitation conditions and a third letter related to air temperature.

The updated version, valid for the second half of the 21st century, is based on monthly, global grids of temperature data from CRU (CRU TS 2.1) (Mitchell and Jones, 2005) and observation based precipitation from GPCC (VASClimO v.1.1) (Beck et al., 2005), both covering the period 1951 – 2000. The map is provided in 0.5° x 0.5° spatial resolution and shows the occurrence of a total number of 30 climate zones (Kottek et al., 2006).

3.2.4 Data Preparation

The temperature data is converted from Kelvin to Degree Celsius and net radiation from Watt per square meter to Mega Joule per square meter and day. Latent energy is converted from Mega Joule per square meter and day to evapotranspiration in millimeter per day using a constant latent heat of vaporization of 2.45 MJ m^{-2}. Further, negative values are set to zero.
The GRACE TWS values must be scaled according to the local scaling coefficients for land area. Afterwards, the dataset as well as the associated uncertainty layer are resampled to a 1° x 1° lat/lon grid by applying the average of the sub-grid cells. Further, the data values are converted from centimeter to millimeter. As June 2003 is missing in the GRACE product, NoData is entered for the corresponding time step.
The GlobSnow SWE dataset as well as the BFI1-4 and K products are aggregated by applying the average of the sub-grid cells to receive the intended spatial resolution. Negative values that occur in BFI2 and BFI4 are set to zero to avoid negative modelled fluxes.
In order to reduce the number of classes, but still represent spatial heterogeneity, Köppen-Geiger climate zones are aggregated primarily based on the main climate. However, as the snow climate covers 56 % of the study area, it is split depending on humidity and summer temperatures to the temperate, boreal and boreal semi-arid zones as illustrated in Fig. 11. Further, the map is resampled by applying the majority class to each 1° grid.
Data Preparation for all datasets includes extraction of the relevant time period and the delineated study area.

Due to the differing temporal coverage of the input datasets, only the overlapping period from 01.01.2003 to 31.12.2010 is considered. Overlap also exists from April 2002 to December 2002 but goes along with two missing months of GRACE TWS data. Thus, the study focusses on the period 2003 – 2010.

As the focus lies on snow-affected regions of the Northern Hemisphere, the study area is delineated by using the GlobSnow SWE data to define a minimum SWE of 30 mm that has to be recorded once in the period 2003 – 2010. In order to minimize the effects of land/water mixed pixels, the water mask is used to extract grid cells with more than 50 % land area. Further, meteorological forcing data must be available to run the model, so pixels without data for precipitation, temperature, net radiation, and evapotranspiration are excluded. The remaining area can be potentially used to run the model. However, for optimization, GRACE TWS and GlobSnow SWE data need to be available, too. Therefore, only cells with one missing GRACE TWS month (i.e. June 2003), measured TWS variations (anomaly ≠ 0) and GlobSnow SWE data for at least 10 % of all days are considered to be valid for optimization. Applying these

constraints, the total amount of 7879 grid cells is reduced by 21.4 % to 6190 valid cells (Fig. 12). The resulting study area is depicted in Fig. 13.

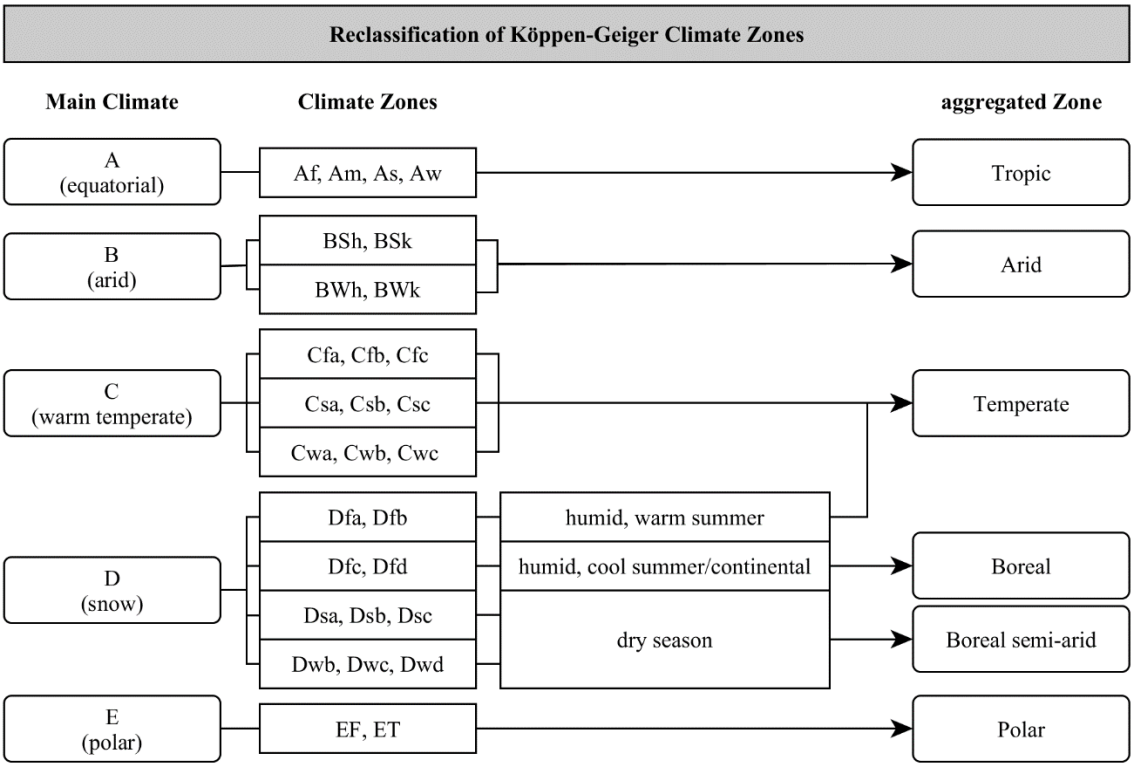


Fig. 11: Aggregation of Köppen-Geiger climate zones. (Precipitation: W – desert, S – steppe, f – fully humid, s – summer dry, w – winter dry, m – monsoonal; Temperature: h – hot arid, k – cold arid, a – hot summer, b – warm summer, c- cool summer, d – extremely continental, F – polar frost, T – tundra) (Kottek et al., 2006).

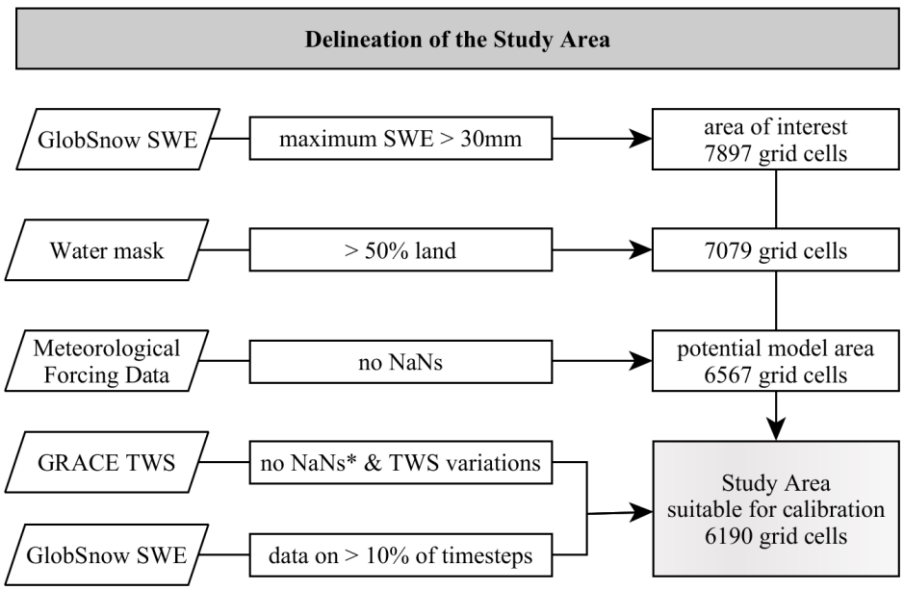


Fig. 12: Delineation of the study area (not more than one NaN due to missing measurements in June 2003).*

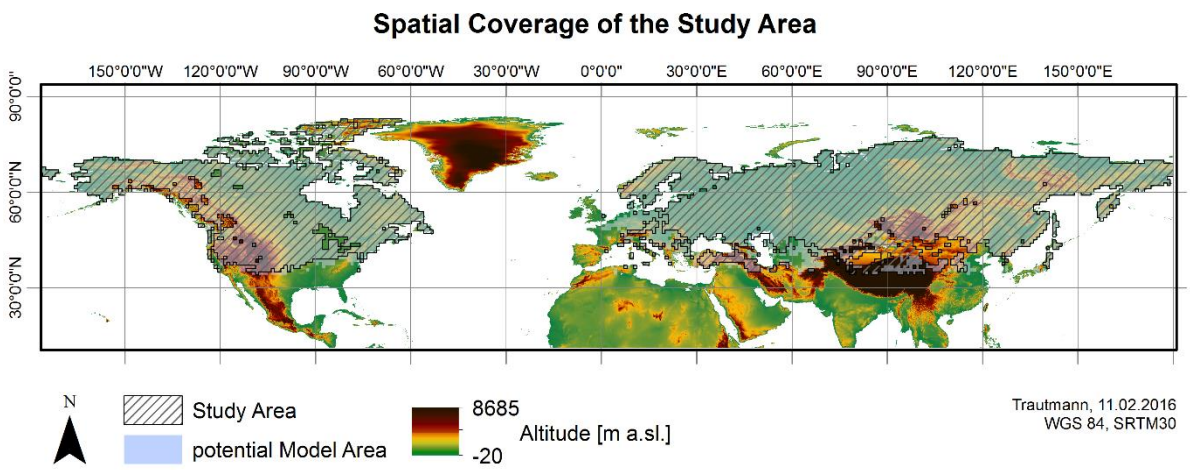


Fig. 13: Spatial coverage of the study area and potential model area.

3.3 Experimental Design

Fig. 15 shows the detailed procedure of this study, which consists of six parts:

1) Comparison of optimization methods
2) First investigation and model ranking
3) Incorporation of spatial distributed information
4) Final model calibration
5) Model evaluation and ranking
6) Application and analysis of the best model variant

The general structure and interplay of these parts is explained in the following, while the individual methodologies and settings are described more precisely in subsequent sections. Tab. 6 provides an overview on the time series considered for the individual steps.

In the beginning, all seven model variants are initialized with globally constant default parameter values based on literature. These parameters shall be optimized using automatic calibration to improve the fit of the model simulations to the observed GRACE TWS, GlobSnow SWE and ET data. Instead of optimizing the parameter set itself, a vector $X = [x_1, x_2 \ldots x_n]$ consisting of n scaling factors x, each associated with one parameter of the model under consideration, is optimized. Accordingly, upper and lower bounds of each parameter are scaled by its default value, when used to constrain the search algorithm. Since GRACE TWS is available only in monthly time steps, the monthly average of the daily SWE, TWS, and ET data are applied for optimization. Furthermore, the models are run for two (basic runoff

variants) resp. 20 (groundwater variants) years prior to estimating the costs in order to initialize water storages. Thereto, years with available forcing data are mixed randomly to prevent reflection of potential trends. Optimization intends to find the parameter set that achieves the best fit between simulations and observations for all grid cells simultaneously. Thus, a global uniform parameter set is determined for each model variant, rather than several, spatial varying parameter sets.

1) Comparison of optimization methods

To ascertain the most suitable optimization method that finds the global optimal parameter set in a minimum of computational time, while being able to account for known data issues, different search algorithms and cost functions are tested and evaluated. In the first step, four commonly applied search algorithms are used to optimize the initial cost function CF1 for all model variants for 1000 randomly chosen grid cells (Fig. 14b). Based on computational time and comparison of the results, one search algorithm is chosen and used for all subsequent procedures.

In the next step, the performance of six different cost functions and their ability to account for data uncertainty, especially regarding the GlobSnow SWE product, is assessed. To reduce the computation time, the cost functions are optimized for 100 randomly chosen pixel (Fig. 14a). The combination of the most suitable cost function and search algorithm represents the 'best' optimization method, that is applied for final model calibration. More details on the comparison of optimization methods is given in chapter 3.3.2.

2) First investigation and model ranking

Before final calibration, a first investigation of the model variants optimized for 100 grid cells is conducted. This investigation includes a ranking of the model variants as well as analysis of the corresponding parameter values. The methodologies are described in chapter 3.3.3. Thereby analysis revealed that optimized parameter values largely differ for the common snow component (chapter 5.1.3). As one aim of this thesis is to identify differences in model performance related to the runoff generation processes, disparity in the snow component may distort the effect of varying soil and runoff components on soil moisture and runoff simulations. Therefor the snow parameter set of the best ranked model is chosen and fixed for all variants during final calibration.

3) Incorporation of spatial distributed information

To examine whether spatialization of model parameters and thus consideration of spatial heterogeneity improves model performance compared to global uniform physio-geographic characterization, the best ranked groundwater variant is selected to implement exemplarily

spatial distributed information on base flow characteristics (Beck-GW variants). The implementation is illustrated in chapter 3.3.4.

4) Final model calibration

Eventually, the soil and runoff parameters of all model variants, including the Beck-GW models, are calibrated for 1000 grid cells using the best optimization method (chapter 3.3.5).

5) Model evaluation

Model performance is evaluated by applying the optimized parameter values for the 5190 grid cells not used for calibration. Additionally, all standard model variants are run for the entire study area (global run). Their costs for calibration, evaluation and global run are evaluated and their simulated global mean seasonal cycles are contrasted. Beck-GW models are compared with their standard counterpart based on their calibration results, and only the best performing Beck-GW is applied globally and considered when analyzing the mean seasonal cycle. More details are given in chapter 3.3.6.

6) Model application

Finally, the 'best' model variant is chosen based on its rank and its simulations are analyzed concerning the research questions as described in chapter 3.3.7.

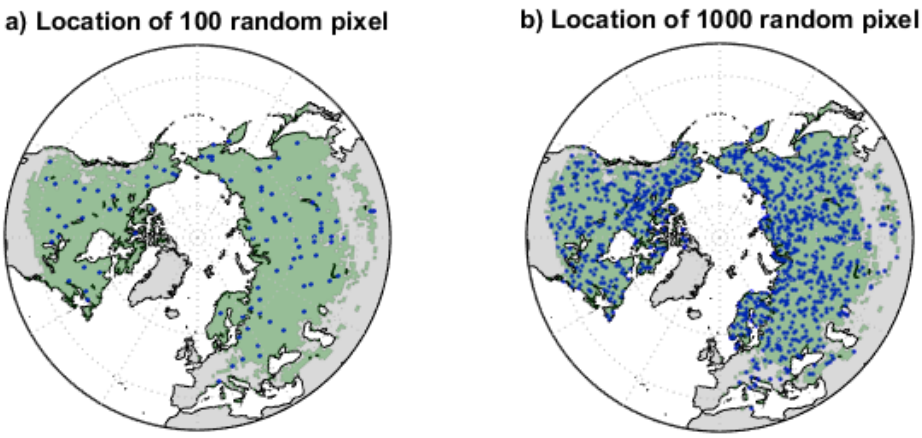


Fig. 14: Location of the a) 100 grid cells used for comparing cost functions and b) 1000 grid cells used for model optimization. Green shaded area marks the study area.

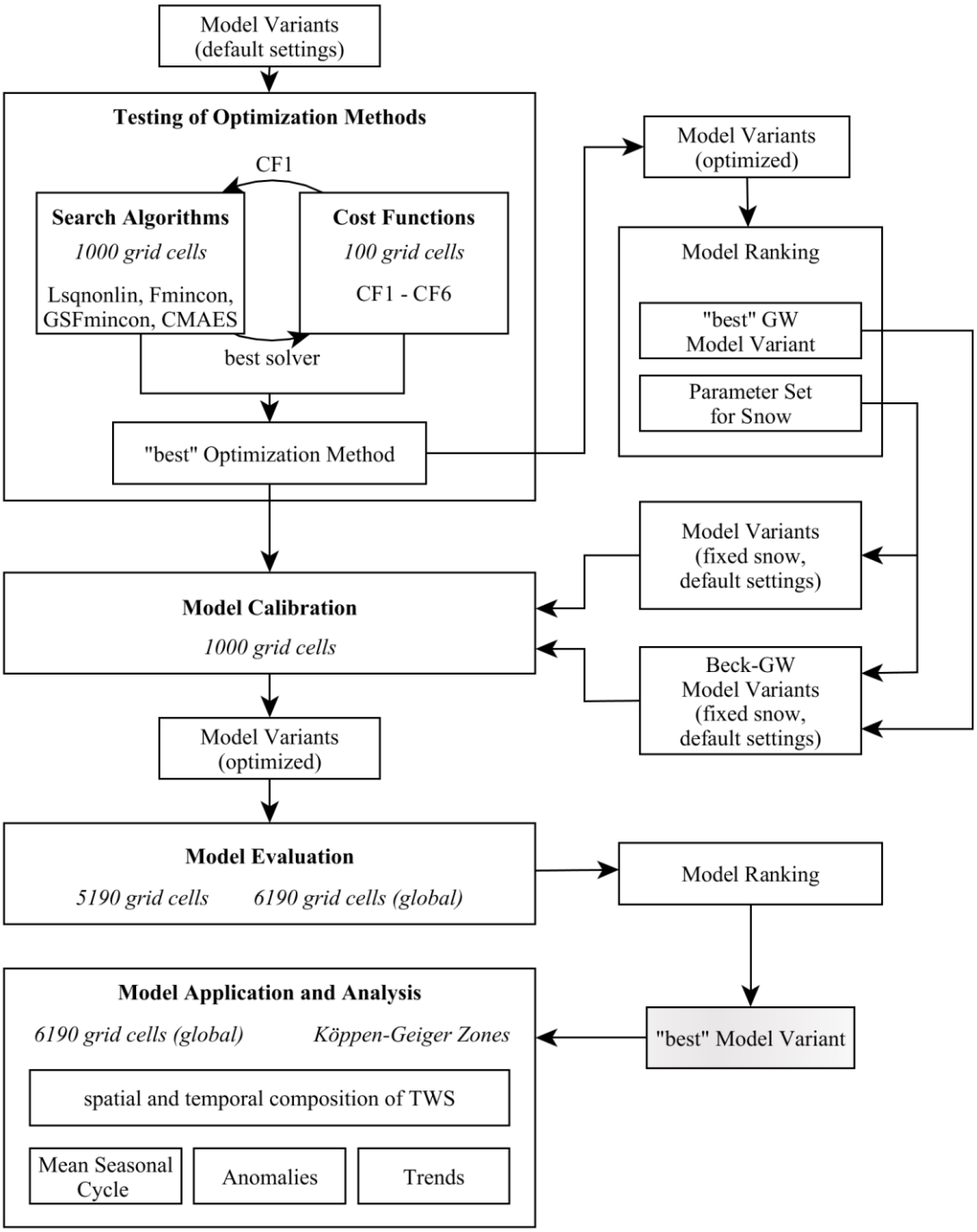


Fig. 15: Experimental design of the master thesis.

Tab. 6: Used temporal coverage of the global data sets and time series considered for model run, evaluation of the cost functions during comparison of optimization methods, calibration, evaluation and analysis of the best model's results.

	Considered Time Series
Global data sets	2003 – 2010
Model run	initialization (2/20 years) + 2003 – 2010
Evaluation of cost function during optimization & calibration	2003 – 2010
Model evaluation and analysis	2004 – 2010

3.3.1 Default Parameter Set and Bounds

The default parameter values as well as their upper and lower bounds are listed in Tab. 3.

As by default no scaling of the forcing data streams is assumed, the parameters ***p_sf*** and ***p_et*** are initially set to 1. In theory, they may vary between 0 and ∞, yet an upper bound of 3 is defined for optimization.

Regarding the snow component, typical values of the degree-day factor ***m_t*** are around 3 °C mm^{-1} (Müller-Schmied et al., 2014; Stacke et al., 2011), but may extend to approximately 5 mm $°C^{-1}$ (Julander, 2000). In this study, 10 mm $°C^{-1}$ is defined as the upper reasonable bound for snow melt. With an enthalpy of fusion of ice of 0.3337 MJ mm^{-1}, the physical upper limit of ***m_r*** is 3 mm MJ^{-1} m^2 d^1 and the parameter initially is set to 2 mm MJ^{-1} m^2 d^1. ***sn_α*** serves as a calibration parameter for computing sublimation and according to Miralles et al. (2011), a constant value of ***sn_α*** = 0.95 is by default suitable for every snow covered pixel. For ***sn_c*** the default value of 15 mm is adopted from Balsamo et al. (2009). As upper limit the very seldom occurring amount of 1000 mm SWE is chosen.

Dependent on soil type and root depth, the water holding capacity for a soil of 1 m depth varies between 60 and 590 mm (AG and Bodenkunde, 2005). Although soil depth as well as root depth differ considerably, by default a value of ***s_max*** = 300 mm is assumed, which approximately is the mean of this range.

s_fac_simple by definition varies between 0 and 1, and as default the mean is applied. ***s_exp_simple*** can theoretically take any value between 0 and ∞, yet an upper limit of 20 is defined for optimization. Initially assuming a linear storage, the default is set to 1, so that with ***s_fac_simple*** = 0.5 half of the available water contributes to runoff.

Müller-Schmied et al. (2014) calibrated ***s_exp_berg*** against mean annual river discharge at 1319 gauging stations on a 0.5 x 0.5° grid. Thereto, they considered a range from 0.0001 – 5 and found ***s_exp_berg*** > 1 suitable for most of the catchments. This study adopts the upper and lower bounds of Müller-Schmied et al. (2014) and uses ***s_exp_berg*** = 1.1 as default value.

The Budyko approach has been applied by Zhang et al. (2008) on a daily time scale for 265 catchments in Australia, where the mode of the calibrated ***s_exp_budyko*** lay between 0.5 and 0.66. Thus, an initial value of 0.6 is assumed, while the parameter may range between 0 and 1.

For the basic variant, Orth et al. (2013) report an recession time scale of 2 days, which is used as default value for ***q_t***. However, the parameter may be calibrated between 0 and 100 days, assuming that even in a 1° x 1° grid cell water that is not retained in soil will contribute to runoff within three months.

Regarding the groundwater variants, ***g_r*** and ***g_d*** range from 0 to 1. For ***g_r***, the default value is set to 0.16, which was found as average for 10 catchments in New Jersey by Alley (1984). Concerning ***g_d***, Alley (1984) identified an average value of 0.26 d^{-1}, while in WaterGAP 2.2

a globally uniform value of 0.01 d^{-1} and in WBM_{plus} of 0.0167 d^{-1} is assumed (Müller-Schmied et al., 2014). Considering the global application, 0.01 d^{-1} is adopted as default in this study.

3.3.2 Comparison of Optimization Methods

The following section briefly introduces the settings of the tested search algorithms and explains the differences between the evaluated cost functions.

Search Algorithms

In the course of this thesis, four commonly used optimization techniques are tested, including the gradient-based local methods Lsqnonlin and Fmincon, a global multi start algorithm applying Fmincon (henceforth GSFmincon) and the heuristic CMAES approach by Hansen and Kern (2004). Lsqnonlin is chosen as the least-square methods are most commonly used for optimization problems (Xu, 2002) and Fmincon as it in comparison to Lsqnonlin handles parameter bounds (MathWorks, 2015b). Since these local methods may not provide the best possible parameter set, GSFmincon investigates a greater portion of the parameter space and is more likely to deliver the global optima. Out of stochastic methods, CMAES showed superior performance in finding the best (global) solution (Arsenault et al., 2013) and according to Hansen and Kern (2004) outperforms gradient-based methods as GSFmincon if the response surface is rugged. Thus, CMAES is tested as fourth search algorithm.

The techniques have been explained in detail in Chapter 2.3.2, while Tab. 7 provides a short summary of the methods.

Tab. 7: Overview on tested optimization techniques.

	Lsqnonlin	**Fmincon**	**GSFmincon**	**CMAES**
	local		global	
	non-linear least square fitting	non-linear constrained	multi start applying Fmincon	Covariance Matrix Adaption Evolution Strategy
Code	MATLAB Optimization Toolbox (MathWorks, 2015b)		MATLAB Global Optimization Toolbox (MathWorks, 2015a)	source code v3.61 (Hansen and Kern, 2004)
Algorithm	Levenberg-Marquardt	interior-point	scatter-search, interior-point	evolution strategy with adaption of covariance matrix

For application of the optimizers, an initial starting point, constraints, possible termination criteria and some optimizer-dependent variables need to be set. For testing, the options are defined as shown in Tab. 8. If not listed otherwise, the function's default settings are used.

Stopping criteria as the maximum number of iterations (MaxIter), the maximum number of function evaluations (MaxFunEval) as well as termination tolerances for the cost function (TolFun) and parameter scaling factors (TolX) are common to all optimizers. For CMAES,

MaxIter depends on the population size, which is calculated according to the number of problem dimensions (number of model parameters) and adjusted so that it's likely to obtain a global solution. The initial step size of CMAES (pSigma) is set in dependency on the upper (ub) and lower parameter bounds (lb). DiffMinChange and DiffMaxChange refer to the gradient based methods and serve to compute the initial search direction. GSFmincon additional requires the number of starting points nStart.

The initial starting point for each model is defined by the default parameter values listed in Tab. 3. Optimization is constrained to find solutions within the upper and lower parameter bounds, except for Lsqnonlin which does not handle bound constraints.

To determine the most suitable search algorithm, their computation time as well as their best solution obtained for calibration of 1000 randomly chosen grid cells and applying initial cost function CF1 are compared.

Tab. 8: Optimization settings used for testing the performance of the four optimization methods and for comparing the suitability of various cost functions.

Option	**Description**	**Value**	
		Lsqnonlin, Fmincon, GSFmincon	**CMAES**
PopSize	population size, depending on the number of problem dimensions N	-	$3(4+(3\log N)$
pSigma	initial step size	-	$0.3(ub-lb)$
nStart	number of starting points (GSFmincon)	50	-
MaxIter	maximum number of solver iterations	1000	$e^3\frac{(N+5)^2}{\sqrt{PopSize}}$
MaxFunEval	maximum number of function evaluations	1000	1000
DiffMinChange	minimum change in variables for finite-difference gradients	e^{-2}	-
DiffMaxChange	maximum change in variables for finite-difference gradients	e^{-1}	-
TolFun	termination tolerance on the cost function value	e^{-6}	
TolX	termination tolerance on the scaling factor	e^{-6}	

Cost Functions

In order to account for different model aspects, total costs consist of three components, that are summed up equally. As the main objective is to simulate available moisture adequately, the first component contrasts the overall fit of modeled TWS (TWS_{mod}) with GRACE TWS data (TWS_{obs}). Therefore, the monthly values of both TWS datasets are translated as anomalies to the time-mean baseline of the period 01.01.2003 – 31.12.2010 and TWS costs are estimated for grid cells and time steps with available observations. To minimize the errors in TWS_{mod} due to snow simulation, the second cost component compares simulated SWE (SWE_{mod}) with GlobSnow SWE (SWE_{obs}). Similar to TWS, SWE costs are calculated only if SWE_{obs} data is

present. As SWE_{obs} is subject to saturation effects (see chapter 3.2.2), the aim is to incorporate the increasing uncertainty related to high SWE_{obs} values. Finally, the third cost is implemented to punish differences between modeled (ET_{mod}) and the observation based ET (ET_{obs}) in order to diminish inconsistencies of model output and forcing data streams. Those may occur if the forcing data suggests $ET_{obs} > 0$, yet the modelled soil water storage is empty and thus $ET_{mod} = 0$ is simulated.

In the context of this study, six distinct cost functions are developed and tested for their suitability. An overview is provided in Tab. 9.

Tab. 9: Cost functions tested within this study. The table shows according to which formula the cost components are calculated, which estimate of uncertainty σ is used and whether the data is further trimmed.

Cost	TWS cost		SWE cost		ET cost		Trimming
Function	**formula**	**σ**	**formula**	**σ**	**formula**	**σ**	
CF1	KGE	-	ρ	-	KGE	-	-
CF2	KGE	-	KGE	-	KGE	-	-
CF3	wMSE	σ_{TWS}	wMSEa	σ_{SWE}	wMSE	σ_{ET}	✓
CF4	wMEF	σ_{TWS}	wMEFa	σ_{SWE}	-	-	✓
CF5	wMEF	σ_{TWS}	wMEFa	σ_{SWE}	wMEF	σ_{ET}	✓
CF6	wMEF	σ_{TWS}	wMEFb	σ_{SWE}	wMEF	σ_{ET}	✓

The first cost function (CF1) utilizes the ED of the KGE for the TWS and ET cost components to evaluate the overall fit of modelled and observed data streams (Eq. (10)). ED is chosen, since, in comparison to KGE, it is subject to minimization and thus directly applicable as cost. Regarding TWS, the bias as defined for the KGE equals unity, because TWS_{mod} and TWS_{obs} represent anomalies to the time-mean base line. Therefore, this component does not affect the ED of TWS cost and is neglected. Furthermore, CF1 attempts to account for increasing uncertainty related to high SWE_{obs} values by trying to match SWE_{obs}'s seasonal variation rather than its magnitude. Thereto, CF1 calculates the SWE cost component using Spearman's rank correlation coefficient ρ instead of ED, and thus neglects the bias between SWE_{obs} and SWE_{mod}. For costs, ρ is subtracted from unity so that high correlation is associated with a lower cost value.

Unlike CF1, the second cost function (CF2) evaluates the overall fit between observed and simulated SWE and applies the ED of the KGE for all three cost components.

While cost function CF1 and CF2 are solely based on the observation data streams, cost functions CF3 to CF6 consider the uncertainty σ of the observation data directly.

Thereto, the third cost function (CF3) calculates the mean square error weighted with the uncertainty of the observational data for each cost component (Eq. (3)).

In a similar way, cost function 4 (CF4) uses the weighted model efficiency for TWS and SWE to evaluate the overall model fit. Unlike the original formula of wMEF (Eq. (8)), the ratio of weighted squared residuals to weighted squared variance is not subtracted from unity, so that the cost is subject to minimization. In order to let the optimizer scale ET within the bounds of ***p_et***, CF 4 does not include a cost component for ET.
Cost function 5 (CF5) and 6 (CF6) consider ET costs again and calculate all three cost components based on wMEF, but differ in the way they account for the uncertainty of SWE.

For CF3-6 the uncertainty of TWS (σ_{TWS}) is based on the uncertainty layer provided with GRACE mascon data (Fig. 16c).
Regarding ET_{obs} a relative uncertainty of 10 % is assumed. In order to impede division through zero (Eq. (3), (8)), the minimum uncertainty is set to 0.1 mm. The resulting applied uncertainty σ_{ET} as a function of ET_{obs} (Eq. (45)) is shown in Fig. 16a.

$$\sigma_{ET} = \max(\, 0.1\, ET_{obs}, 0.1) \tag{45}$$

Based on reported differences to ground-measurements (Liu et al., 2014; Luojous et al., 2014), an average of 35 mm absolute uncertainty is applied for SWE_{obs} below a saturation threshold (thrsh) of 100 mm (Fig. 16b). Above this threshold, evaluation of GlobSnow SWE suggests that the dataset provides no reliable information on the actual SWE amount. Thus, the performance of SWE_{mod} is not compared to SWE_{obs} for SWE_{obs} > thrsh. Instead, it is assumed that SWE_{mod} provides good estimates of the actual SWE as long as both, SWE_{obs} and $SWE_{mod,}$ exceed 100 mm. The way how costs are calculated if SWE_{obs} exceeds the threshold differs among wMSEa resp. wMEFa and wMEFb. In wMSEa /wMEFa costs are calculated for the difference between SWE_{mod} and the threshold, if SWE_{mod} is less than the thrsh but SWE_{obs} exceeds it. If both, SWE_{mod} and SWE_{obs}, exceed 100 mm, costs are set to zero (Eq. (47) – Eq. (50)).

$$\sigma_{SWE} = 35 \tag{46}$$

$$SWEa(SWE_{obs} > thrsh) = \frac{\max(\, trsh - SWE_{mod}(SWE_{obs} > thrsh), 0)^2}{{\sigma_{SWE}}^2} \tag{47}$$

$$SWEa(SWE_{obs} \leq thrsh) = \frac{(SWE_{obs}(SWE_{obs} \leq thrsh) - SWE_{mod}(SWE_{obs} \leq thrsh))^2}{{\sigma_{SWE}}^2} \tag{48}$$

$$wMSEa = \frac{1}{N}\sum SWEa \tag{49}$$

$$wMEFa = \frac{\sum SWEa}{\sum \frac{(SWE_{obs} - \mu_{SWEobs})^2}{\sigma_{SWE}{}^2}} \tag{50}$$

For wMEFb, values of SWE_{obs} and SWE_{mod} that exceed the threshold are set to the threshold before calculating the costs (Eq. (51) – Eq.(53)).

$$SWEb_{obs}(SWE_{obs} > thrsh) = thrsh \tag{51}$$

$$SWEb_{mod}(SWE_{mod} > thrsh) = thrsh \tag{52}$$

$$wMEFb = \frac{\sum \frac{(SWEb_{obs} - SWEb_{mod})^2}{\sigma_{SWE}{}^2}}{\sum \frac{(SWEb_{obs} - \mu_{SWEbobs})^2}{\sigma_{SWE}{}^2}} \tag{53}$$

Additionally, in contrast to CF1 and CF2, CF3-6 further trim the data to exclude outliers in cost calculation. Thereto, the 95th percentile of the residuals TWS_{obs} – TWS_{mod} resp. SWE_{obs} – SWE_{mod} is calculated and used as a threshold to exclude data points with a greater residual.

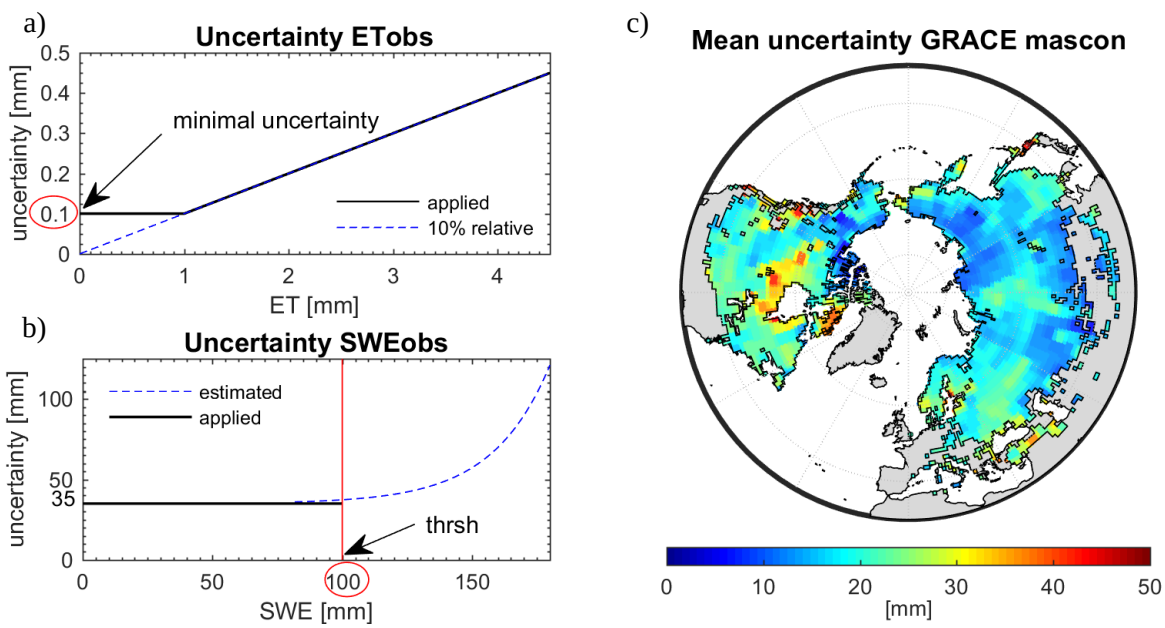


Fig. 16: Applied uncertainty [mm] as a function of a) ET_{obs} and b) SWE_{obs} as well as c) spatial distribution of mean TWS_{obs} uncertainty [mm].

3.3.3 First Investigation and Model Ranking

For first investigations and to determine the best groundwater model variant, the best solutions obtained with the most suitable search algorithm and cost function for 100 grid cells are inspected. Therefor the cost function values are examined and used to rank the models. According to Gulden et al. (2008), the mean rank (mean rk) is calculated as the average of the ranks (rk) for TWS, SWE and ET costs (Eq. (54)). To consider the gap between model costs in relation to differing ranges of the cost components, the weighted mean rank (w mean rk) is computed by weighting the ranks of the cost components with their corresponding range (R) (Eq. (55)).

$$mean\ rk = \frac{rk_{TWScost} + rk_{SWEcost} + rk_{ETcost}}{3} \tag{54}$$

$$w\ mean\ rk = \frac{(rk_{TWScost}\ R_{TWScost}100) + (rk_{SWEcost}R_{SWEcost}100) + (rk_{ETcost}\ R_{ETcost}100)}{3} \tag{55}$$

3.3.4 Incorporation of spatial distributed Base flow Information

To test whether model performance can be further improved by considering spatial distributed information, BFI and K estimates by Beck et al. (2015) are implemented. These Beck-GW models are identical to the best groundwater model variant of the first investigation, yet calculate runoff by applying the spatial distributed information of either K (ModelGWBeck-K), BFI (ModelGWBeck-BFI1-4), or both, BFI and K (ModelGWBeck1-4) (Fig. 17). Since Beck et al. (2015) provides four BFI estimates, the combination results in nine different Beck-GW models.

The recession coefficient K [d^{-1}] corresponds to the parameter ***g_d*** [d^{-1}] of the standard groundwater variants and substitutes this parameter when calculating Qbase [mm d^{-1}] in *CalcRunoffGW_K* and *CalcRunoffGW_Beck*. BFI is defined as the ratio of the long-term base flow to total streamflow and thus relates to the percolating fraction of surface runoff Qs. Hence, it replaces the parameter ***g_r*** [-] in the MATLAB functions *CalcRunoffGW_BFI* and *CalcRunoffGW_Beck*.

Further, K as well as BFI are allowed to be scaled by the parameters ***p_gd*** [-] resp. ***p_gr*** [-] that both are by default set to 1, but can be adjusted during final calibration between a lower bound of 0 and an upper bound of 3 similar to ***p_sf*** and ***p_et***. To ensure that the percolating fraction of Qs and the fraction of groundwater storage GW that forms base flow do not exceed 100 %, the minimum of ***p_gr***·BFI and 1 is applied to calculate percolation Perc [mm d^{-1}] and the minimum of ***p_gd***·K and 1 to derive base flow Qbase [mm d^{-1}].

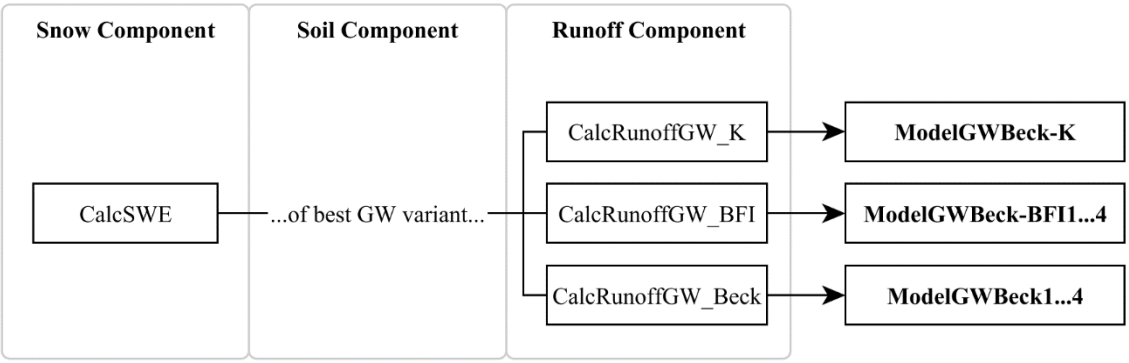


Fig. 17: Model configuration of Beck-GW variants.(Calc indicates a MATLAB function, while Model refers to the final combination of these functions).

3.3.5 Final Model Calibration

Final model calibration applies the so found most suitable optimization method to find the optimal global parameter set for 1000 grid cells, for each model variant.

Therefor the optimization settings are modified by increasing the maximum number of function evaluations to 10000 in order to ensure that the algorithm only stops if no further improvement is possible.

To establish confidence in the optimization results, the output of the optimization algorithm with information on the progress of costs and parameters with each iteration is investigated.

3.3.6 Model Evaluation

Model verification includes analysis of the optimized parameters and the evaluation of model performance for independent data. Thereto, the model variants are applied solely for the 5190 grid cells not used for calibration as well as globally for the entire study area. Model costs are compared for the calibration, the evaluation and the global run and used to rank the model variants as described in chapter 3.3.3. Additionally, the mean seasonal cycle (MSC) of all grid cells is investigated to examine whether the model variants differ in their temporal dynamic. To check for spatial divergences in the models' performance, pixel wise costs are calculated and mapped. Additionally, grid cells with costs higher than the 95th percentile and lower than the 5th percentile are identified as areas of especially poor and good model performance.

3.3.7 Model Application

To finally address the research questions, the best ranked model variant is applied for the entire study area and its results are analyzed for the period 01.01.2004 – 31.12.2010.

Analysis includes investigation of the MSC to assess seasonal variations and the anomalies to the mean seasonal cycle to evaluate inter-annual variability. Areas of poor model performance are identified by calculating efficiency criteria as the RMSE, KGE and r additional to the

applied cost function for each grid cell and for each modified Köppen-Geiger (KG) zone. Shifts in the MSC between TWS_{obs} and TWS_{mod} are measured using cross-correlation.
For determination of the composition of TWS, the variability of TWS can be explained by the variability of its components, mainly represented by SWE, SM and GW. With combining SM and GW to total available water W, the variance in TWS (varTWS) results from the sum of the variance in SWE (varSWE), the variance in W (varW) and the covariance of SWE and W (covar(SWE, W)):

$$varTWS = varSWE + varW + 2covar(SWE, W) \tag{56}$$

If SWE and W do not correlate, covar(SWE,W) approximates zero, and varTWS equals to the sum of varSWE and varW (Lehn and Wegmann, 2006).
In a similar way, the variability of W can be expressed by the variability of SM and GW:

$$varW = varSM + varGW + 2covar(SM, GW) \tag{57}$$

Applying Eq. (56) and Eq. (57) for each grid cell on the MSC and the anomaly to the MSC, spatial information on the intra- and inter-annual influence of SWE, SM and GW on TWS is obtained. The main component for each grid cell is determined by calculating the difference between varW and varSWE, both normalized with varTWS:

$$\frac{varW}{varTWS} - \frac{varSWE}{varTWS} \tag{58}$$

Accordingly, positive values indicate a predominant influence of W, while negative values represent an increased impact of SWE.
In the same way, varW is attributed to be more dominated by varSM or varGW.

Additionally, to examine whether a trend in the GRACE TWS data is present and if this is captured by the model, the rank-based non-parametric Mann-Kendall test (Kendall, 1975; Mann, 1945) utilizing Sen's slope for trend estimation (Sen, 1968) is applied for the observed and modelled TWS anomalies. Thereby, trends are considered significant for p values < 0.05.

4 Study Area

The following chapter briefly characterizes the study area concerning the applied observational data for model forcing, calibration and validation. It further introduces the aggregated Köppen-Geiger zones used for analysis of the model output.

Precipitation, Temperature and Evapotranspiration

Fig. 18 illustrates the spatial distribution of mean annual precipitation P [mm a^{-1}] and average air temperature T [°C], while Fig. 19 shows the mean annual evapotranspiration ET [mm a^{-1}], for each grid cell as well as the global mean seasonal cycle of P [mm $month^{-1}$] (blue line) and T [°C] (red line) for the study area and the period 2003 – 2010.

According to the applied meteorological data, the average annual precipitation is 578 mm a^{-1}, of which 291 mm a^{-1} evapotranspire. Thereby P, and simultaneously ET, decrease towards the more continental parts of North America and Eurasia. Lowest P is observed in polar regions as well as in the North American Middle West and from Kazakhstan eastwards to Mongolia. Average T decreases north-eastwards, reaching minima on the Canadian Arctic Archipelago as well as in Middle and East Siberia.

The average MSC of T in the study area (Fig. 19c) is characterized by lowest temperatures in January and highest in July, with an amplitude of ca. 35 K. P in general is lowest in winter and spring, with minimum amounts measured in February. Maximum monthly P is received in July and August, and decreases gradually during autumn.

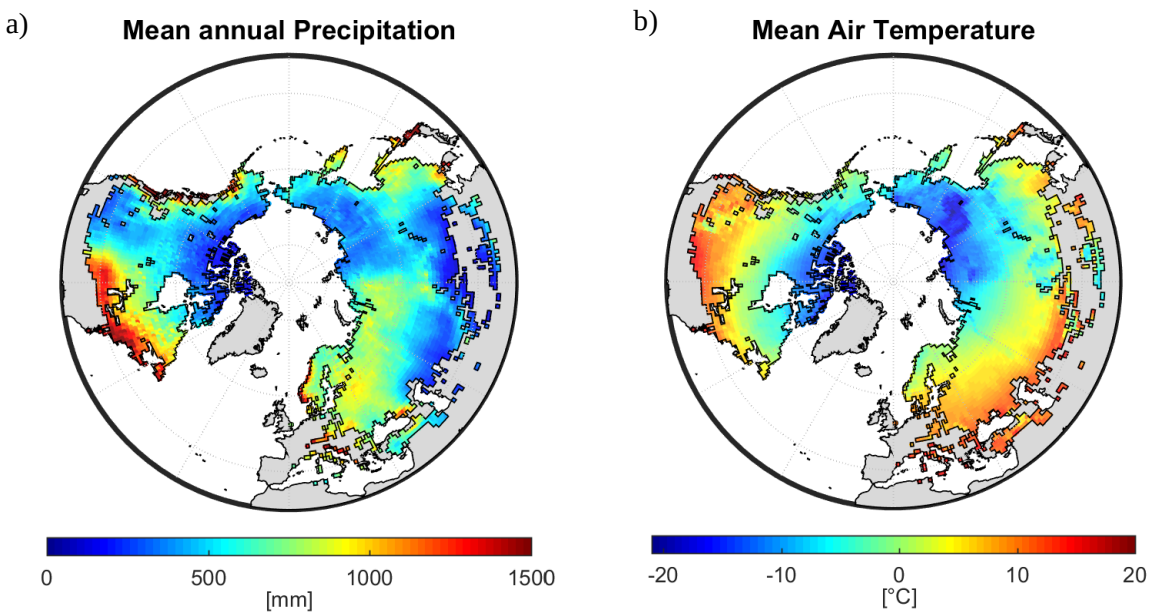


Fig. 18: a) Mean annual precipitation [mm a^{-1}] and b) average temperature [°C] of the study area obtained from the meteorological forcing data for the period 2003 – 2010.

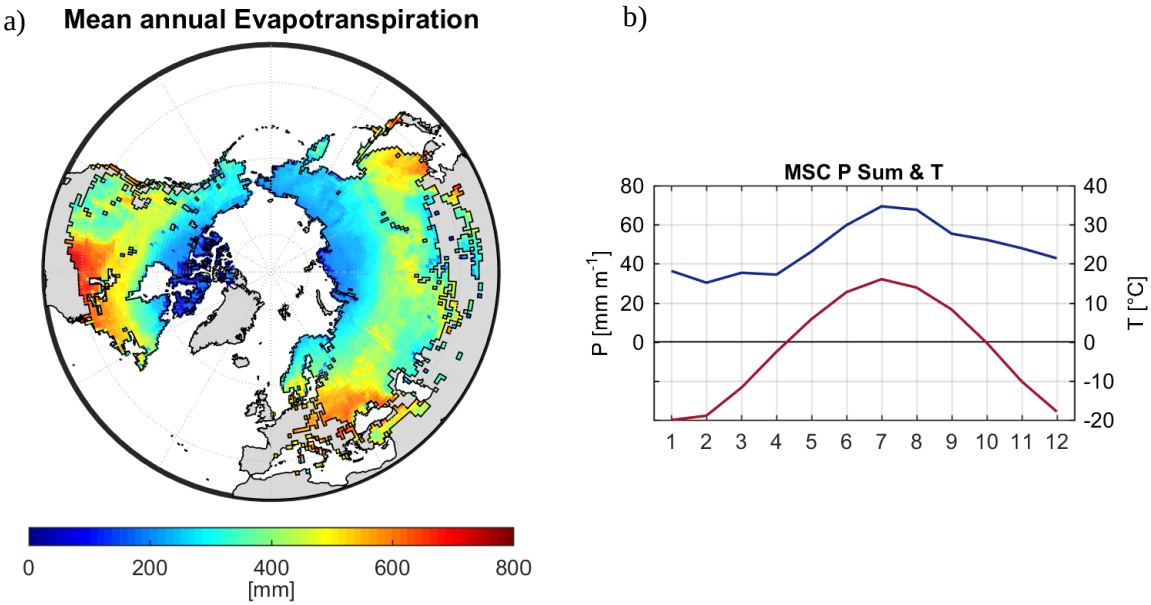


Fig. 19: a) Mean annual evapotranspiration [mm a^{-1}] of the study area as well as b) average mean seasonal cycle of precipitation [mm month^{-1}] and temperature [°C] obtained from the meteorological forcing data for the period 2003 – 2010.

GRACE TWS and GlobSnow SWE

Fig. 20a shows the mean seasonal amplitude of terrestrial water storage variations observed by GRACE (TWS_{obs}) [mm] for each grid cell. The average amplitude of the entire study area is 110 mm, with high variations being present in the north east and along the west coast of North America, the Canadian Arctic Archipelago, North East Siberia including Kamchatka, the West Siberian and East European Plain and Anatolia. Areas of low variations are located in central North America and the South East of Eurasia. On average, TWS_{obs} is at the maximum in April and lowest during September and October.

In Fig. 20b the average of the annual maximum snow water equivalent of the years 2003 – 2010 provided by the GlobSnow product (SWE_{obs}) [mm] is depicted for each grid cell. In general, SWE_{obs} increases from South to North, whereby highest amounts are observed on the Labrador Peninsula, the Canadian Arctic Archipelago, Kamchatka, Middle Siberia and the Scandinavian Mountains. These areas coincidence with medium to high annual precipitation (> 900 mm a^{-1}) and average temperatures below 5 °C (Fig. 18). The average annual maximum of the entire study area is 120 mm, which is observed in March. In spring, SWE_{obs} decreases until June, and snow accumulation starts again in September. However, for July and August the GlobSnow product does not provide SWE data.

Köppen-Geiger Zones

Fig. 21 provides a map of the aggregated Köppen-Geiger (KG) zones and their proportion of the study area. The total number of 30 climate zones is reduced to six classes of which five

occur in the study area. Each of them covers at least 8 % of the area. Most of the grid cells (46 %) are assigned to the boreal zone, followed by the temperate class (25 %), while the polar zone represents 12 % of grid cells. Both, arid and boreal semi-arid zone cover approximately 8 % of the area.

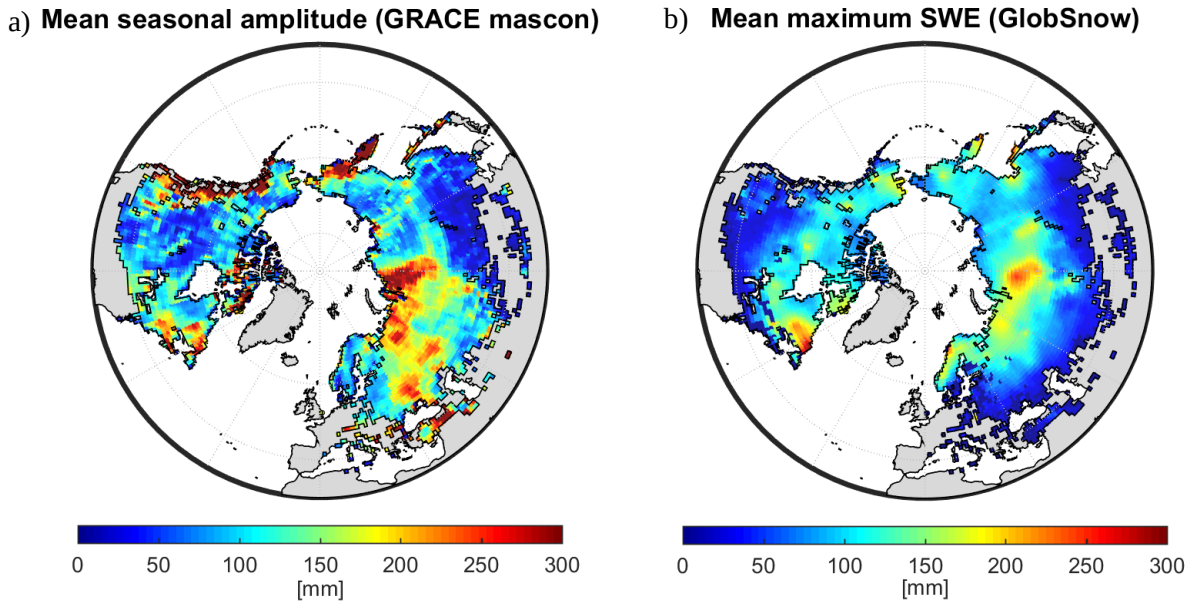


Fig. 20: Mean seasonal amplitude of TWS_{obs} [mm] and mean of maximum annual SWE_{obs} [mm] obtained from the calibration and validation data for the period 2003 – 2010.

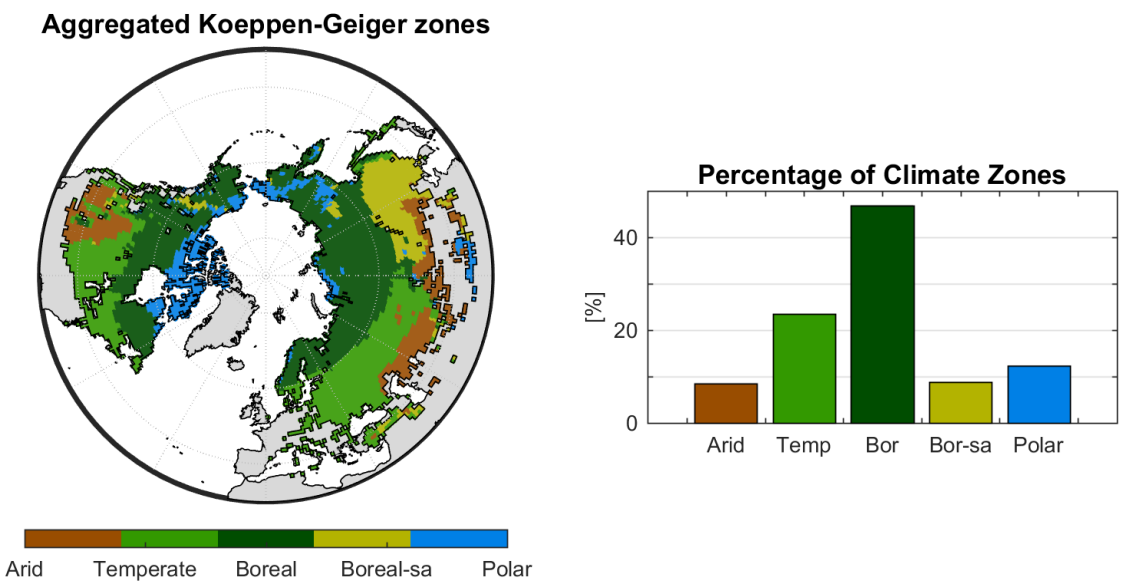


Fig. 21: Spatial distribution of aggregated Köppen-Geiger zones and the corresponding proportion of the study area.

5 Results

The following chapter presents the findings of this master thesis.
In the first part, the results of comparison of optimization methods are shown. Based on these, the most suitable method is chosen for final model calibration, which outcome is shown in the second section of this chapter. The third section deals with the evaluation of the calibrated models and their comparison to infer the most suitable approach. Finally, the last section provides the results of application of this model variant, which are investigated regarding their conformity with TWS_{obs} and the spatial and temporal composition of TWS_{mod}.

5.1 Comparison of Optimization Methods

As one aim of this study is to identify the most suitable optimization method, the following section displays the findings of testing different search algorithm and cost functions. Subsequently, a first model ranking based on these results is conducted.

5.1.1 Search Algorithms

Since the search algorithm shall be least time-consuming, yet provide optimal results, the four search algorithms Lsqnonlin, Fmincon, GSFmincon and CMAES are investigated regarding their run time as well as their obtained solutions.

Run Time

Fig. 22 contrasts the run time needed by Lsqnonlin, Fmincon, CMAES and GSFmincon to optimize each model variant for 1000 grid cells. While the local solvers finish within two hours, GSFmincon needs longest, with on average 3 days and even 6.5 days for BudykoGW. On contrary, CMAES accomplishes on average within 18 hours. In general, optimization of the groundwater variants takes longer than of their basic runoff counterparts.

Comparison of Results (CF1)

The costs (CF1) associated with the best solution found by each search algorithm are depicted for each model variant in Fig. 23. Overall, optimization always reduces costs compared to the models' default parametrization. Thereby Lsqnonlin achieves the least reduction, while Fmincon as another local algorithm obtains better results (except for BergGW), that mostly are comparable to the global search GSFmincon. The latter solely performs substantially better than Fmincon regarding SimpleBasic. Except for BudykoBasic, the largest improvement of costs is achieved with CMAES. However, the difference between Fmincon, GSFmincon and CMAES is mostly less than 0.03, while the optimized parameter values may differ largely (A 3).

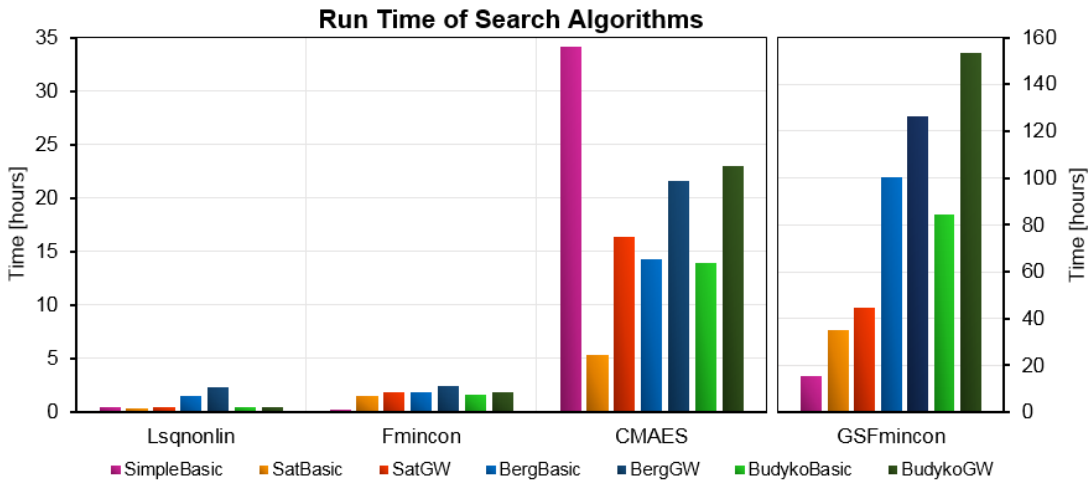


Fig. 22: Run time [hours] of the search algorithm for optimizing each model variant for 1000 grid cells. Algorithm settings as described in Tab. 8.

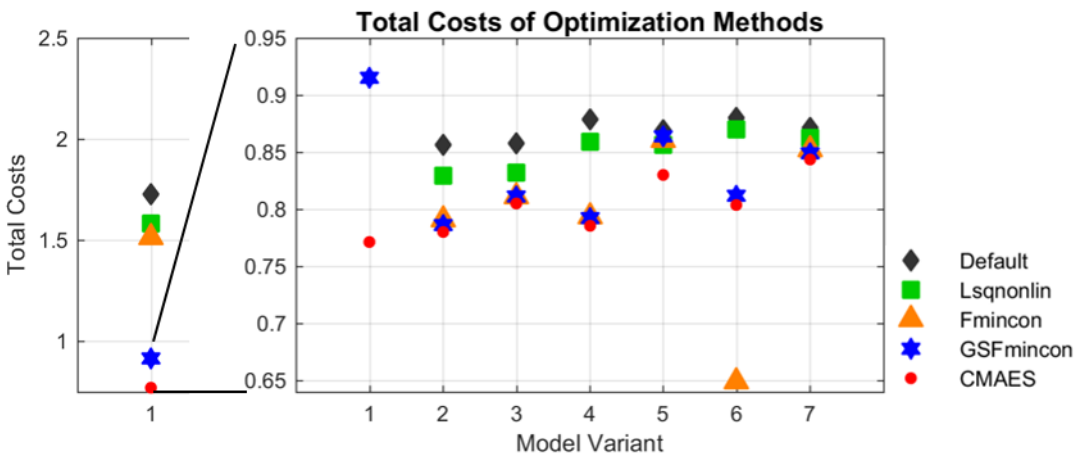


Fig. 23: Total Costs (CF1) for model variants with default parametrization (Default) and optimized parameter values derived by four search algorithms. Model variants: 1-SimpleBasic, 2-SatBasic, 3-SatGW, 4-BergBasic, 5-BergGW, 6-BudykoBasic, 7-BudykoGW; optimization for 1000 pixel regarding CF1.

5.1.2 Cost Functions

One major question when determining the most suitable optimization method is how the cost function shall be formulated to account for known data issues, especially regarding GlobSnow SWE estimates. Therefor six cost functions have been tested, which differ in four issues:

a) how SWE_{obs} uncertainty is treated,
b) whether the data streams are trimmed or not,
c) whether ET costs are considered, and
d) which efficiency criteria is applied.

In the following, the findings concerning these topics are shown.

a) Treatment of SWE_{obs} Uncertainty

Concerning SWE, cost functions shall capture several properties. First, whether the temporal dynamic is reproduced or not and second whether a bias between SWE_{obs} and SWE_{mod} is present. As for the latter increasing uncertainty of SWE_{obs} above a threshold is known, one can distinguish between three cases how the bias should be treated theoretically. If both SWE_{obs} and SWE_{mod} are below the threshold (<< thrsh), bias should be associated with high costs. On contrary, if both values exceed the threshold (>> thrsh), the dataset is assumed to provide no reliable information on the actual SWE amount and thus a bias can be neglected. However, as a fixed value is set as threshold, yet uncertainty is rather gradual increasing, costs should not be too high if SWE_{obs} is just below the threshold (< thrsh) while SWE_{mod} exceeds it (> thrsh).

All in all, this results in four cases that should be considered when evaluating the cost functions. They are summarized in Tab. 10 and exemplarily illustrated in Fig. 24. Fig. 24 further shows how the particular cost functions deal with each case and lists the corresponding costs.

CF1 applies Spearman's rank correlation ρ and thus compares the ranks of SWE_{obs} and SWE_{mod}. This results in high costs if the temporal dynamic is not matched (case 1) and costs of zero if the dynamic is reproduced, regardless of whether a bias exists or not (case 2-4). With CF2 a shift between SWE_{obs} and SWE_{mod} is associated with high costs, too, and the ED of the KGE additional considers bias. Thereby, increasing uncertainty in SWE_{obs} is not taken into account so that, in the example, highest costs are obtained for case 3. Similar, applying wMSEa resp. wMEFa, CF3-CF5 obtain high costs for temporal shift and consider bias. Since costs are set to zero once SWE_{obs} exceeds the threshold, costs are low in case 3, yet relatively high in case 4 (especially for CF4 and CF5). CF6 uses wMEFb and is identical to CF4 and CF5 in case 1 and 2, but assigns lower cost values in case 4.

Consequently, CF1 considers solely case 1 as intended, while CF2 additional treats case 2 properly. Further, CF3-5 capture case 3 and CF6 finally all 4 cases as they should be (Tab. 10).

Tab. 10: Four cases relevant when evaluating the fit between SWE_{obs} and SWE_{mod} and whether they are considered by the cost function as intended.

Case		**1**	**2**	**3**	**4**
Case description					
Temporal Dynamic		×	✓	✓	✓
Bias if	SWE_{obs}	-	<< thrsh	>> thrsh	< thrsh
	SWE_{mod}	-	<< thrsh	>> thrsh	> thrsh
Correct (✓) or wrong (×) consideration by cost function					
CF1		✓	×	×	×
CF2		✓	✓	×	×
CF3/CF4/CF5		✓	✓	✓	×
CF6		✓	✓	✓	✓

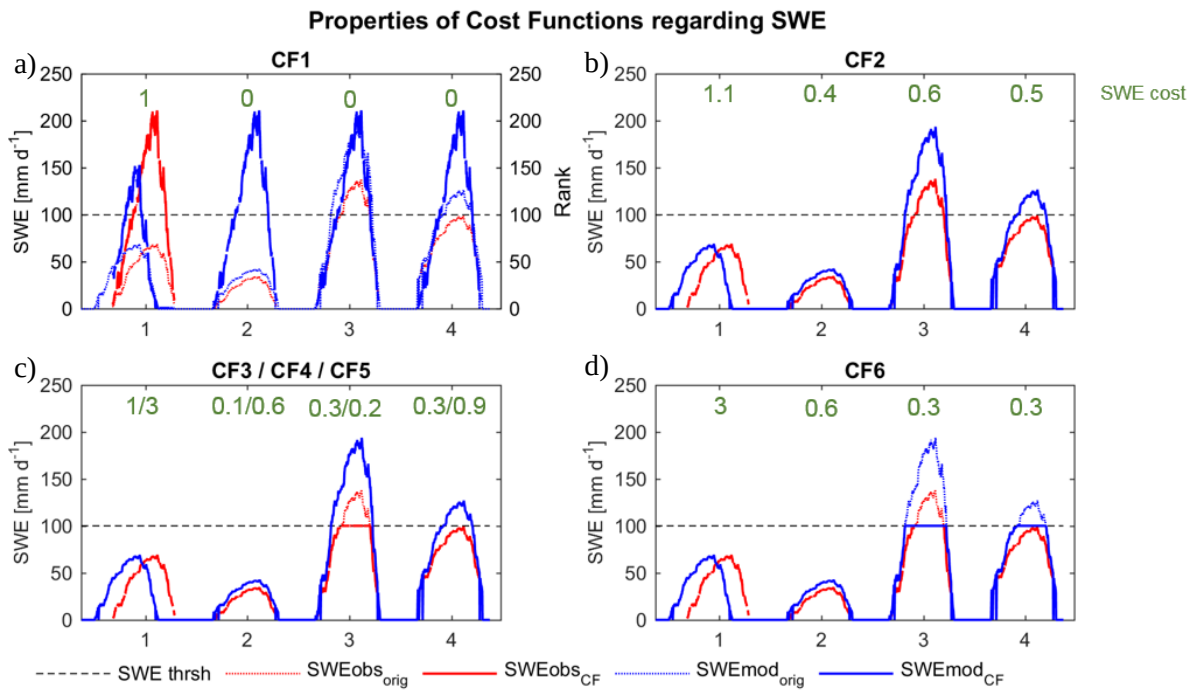


Fig. 24: Properties of the cost functions regarding SWE. Shown are the original time series (dotted lines), assumptions (modifications of SWE_{obs} and SWE_{mod}) made for cost calculation (solid lines) and calculated costs (numbers in green) for the four cases (Tab. 10). In c) the first cost value refers to CF3 while the second number represents SWE costs for CF4 and CF5 (own representation).

b) Trimming of Data Streams

Fig. 25b shows an example of the right skewed distribution of absolute residuals between TWS_{obs} and TWS_{mod} as well as the 95th percentile threshold applied for trimming. As Fig. 25a suggests, differences originate from both, over and underestimation of TWS_{obs}. Thus, while in Fig. 26a extreme values increase the range of mean monthly TWS_{obs} variations, the trimmed time series of TWS_{obs} and TWS_{mod} differ less in their peaks and minima. Correspondingly, costs (CF6) for this example decrease from 0.77 to 0.56.

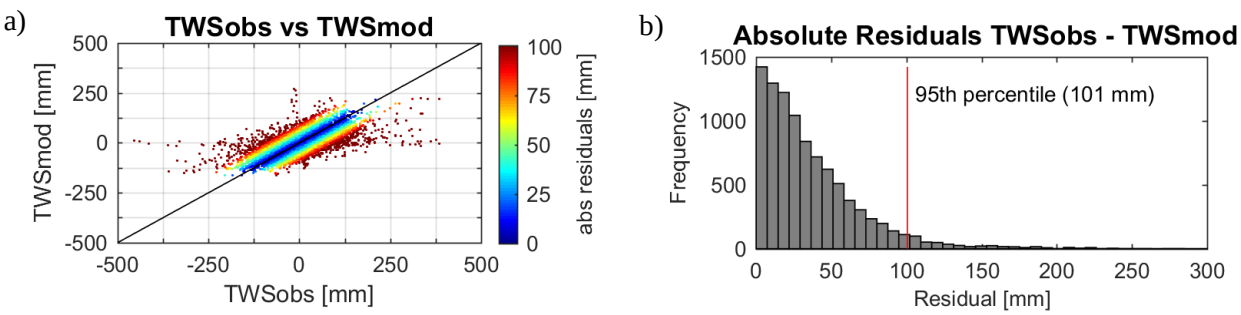


Fig. 25: a) Scatter plot of TWS_{obs} vs. TWS_{mod} and b) histogram of corresponding absolute residuals(SimpleBasic optimized for CF6). In a) the color displays the absolute residuals, with dark red marking the 95th percentile threshold also shown in b).

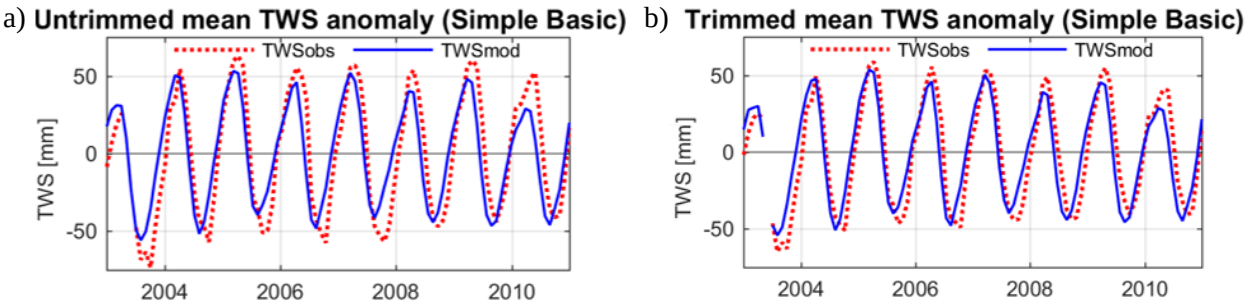


Fig. 26: Mean monthly TWS_{obs} and TWS_{mod} of 100 grid cells. TWS_{mod} is calculated with SimpleBasic optimized for CF6. a) shows the mean of all data points while b) plots the mean excluding data points whose residuals exceed the 95th percentile threshold.

c) Consideration of ET costs

Neglecting ET as a cost component in CF4 results in an extreme reduction of ET_{mod} by modification of the scaling parameter ***p_et*** as shown in Fig. 27. Applying the otherwise identical CF5, reveals ET costs of 1.18 for this example, which increases total costs from 0.9 (CF4) to 2.08 (CF5).

d) wMSE vs. wMEF

Applying wMSE (CF3) instead of wMEF (CF4-6) leads to unrealistic behavior in the seasonal dynamic of SWE_{mod} simulated with some optimized models. For example, Fig. 28 shows, that the mean snow ablation cannot be matched after optimization and that a minimum SWE_{mod} of 24 mm exists throughout the year.

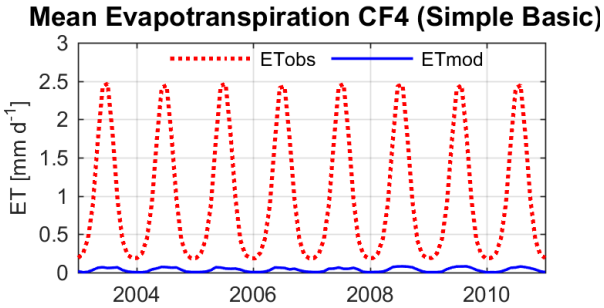


Fig. 27: Mean monthly observed and modelled ET of 100 grid cells. ET_{mod} is calculated with SimpleBasic optimized for CF4.

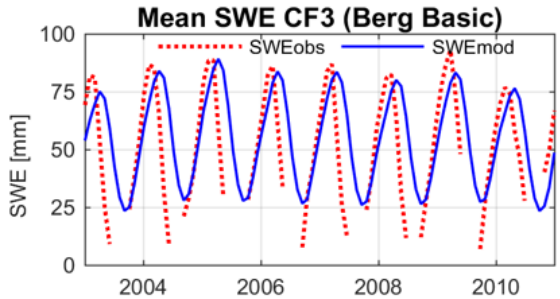


Fig. 28: Mean monthly observed and modelled SWE of 100 grid cells.SWE_{mod} is calculated with Berg Basic optimized for CF3.

5.1.3 First Investigation and Model Ranking

Based on the results obtained by optimization of all model variants for 100 grid cells using CMAES and CF6, a first investigation of the calibrated models is conducted. Therefor they are ranked according to the achieved CF6 value and the optimized parameter values are examined.

Model Ranking

Tab. 11 shows the absolute values of costs (abs) and the corresponding ranking of each model (rk), as well as their averaged rank (mean rk) and their weighted average rank (w mean rk) obtained using CMAES and CF6.

Overall, lowest total costs are scored using the BergBasic variant, closely followed by BergGW, whereas the worst results are obtained by the SatBasic approach. Regarding ET costs, SatGW is considered the best. SWE and TWS costs are lowest for BergGW resp. BergBasic.

The range of costs between the models is extremely low, with 0.07 for total costs, 0.09 for TWS, 0.03 for ET and 0.04 for SWE. As additionally the absolute values of the cost components and thus their influence on total costs differ, the weighted averaged rank seems to be a better index for comparison of overall model fitness than the simple average rank.

However, both, the averaged rank as well as the weighted averaged rank, indicate BergGW to be the best model variant.

This ranking is in general independent from the applied cost functions (A 4).

Tab. 11: Model variant costs based on CMAES and CF6 applied for 100 grid cells,average ranking according to Gulden et al. (2008) and weighted average ranking (abs represents the absolute value of costs and rk identifies the rank among the model variants respectively). Green color highlights the best value resp. rank.

	SimpleBasic		**SatBasic**		**SatGW**		**BergBasic**		**BergGW**		**BudykoBasic**		**BudykoGW**	
Costs	abs	*rk*	abs	*rk*	abs	*rk*	abs	*rk*	abs	*rk*	abs	*rk*	abs	*rk*
Total	1.038	*3*	1.065	*7*	1.045	*4*	**0.996**	***1***	1.003	*2*	1.050	*5*	1.064	*6*
ET	0.047	*3*	0.045	*2*	**0.037**	***1***	0.070	*7*	0.060	*5*	0.061	*6*	0.058	*4*
SWE	0.426	*6*	0.416	*4*	0.412	*3*	0.411	*2*	**0.407**	***1***	0.416	*4*	0.447	*7*
TWS	0.565	*4*	0.604	*7*	0.596	*6*	**0.515**	***1***	0.536	*2*	0.573	*5*	0.558	*3*
mean rk		*4.3*		*3.3*		*4.3*		*3.3*		***2.7***		*5.0*		*4.7*
w mean rk		*28.5*		*23.0*		*23.3*		*13.2*		***12.7***		*26.8*		*22.7*

Parameter Investigation

The optimized parameter values of the snow component obtained using CMAES and CF6 for 100 pixels are listed in Tab. 12. Similar to parameters obtained when testing search algorithms (A 3), they differ considerably among the model variants although all apply the same model formulation and forcing. Solely concerning ***sn_alpha*** all variants correspond by decreasing the parameter, and consequently ETSub, to a minimum. On contrary snow melt is increased by ***m_t*** and ***m_r***, that both (almost) reach their upper bound, except for SimpleBasic and SatGW. Further, ***p_sf*** scales SF to 66 - 84 % of observed P and ***sn_c*** varies between 91 and 492 mm.

As one aim of this thesis is to identify differences in model performance related to the runoff generation processes, disparity in the snow component may distort the effect of varying soil and runoff components on soil moisture and runoff simulations. Therefor the snow parameter set of BergGW is chosen and fixed for all variants during final calibration. This choice is based on the variant's good ranking regarding SWE costs and its overall performance (both, total costs and weighted averaged rank). Additional, the snow parameter set of BergGW reflects roughly the average (***p_sf***, ***sn_c***) resp. corresponds approximately to the majority of optimized values (***m_t***, ***m_r***, ***sn_alpha***), and thus is assumed to be most appropriate to be used as predefined for all model variants.

Tab. 12: Default and optimized snow parameter values applying CMAES and CF6 for 100 grid cells. The BergGW parameter set used for further proceeding are written in bold letters and marked by a green frame.

Parameter		**Default**	**Simple**	**Sat**		**Berg**		**Budyko**	
			Basic	**Basic**	**GW**	**Basic**	**GW**	**Basic**	**GW**
snow	*p_sf*	1.00	0.84	0.75	0.79	0.77	**0.74**	0.76	0.66
	m_t	3.0	6.9	10.0	5.3	10.0	**9.9**	9.8	9.3
	m_r	2.0	1.6	2.8	1.1	2.7	**3.0**	2.8	2.7
	sn_alpha	0.95	0.00	0.01	0.00	0.00	**0.00**	0.00	0.07
	sn_c	15	164	260	91	195	**207**	256	492

5.2 Final Model Calibration

In the following the outcome of final calibration of all model variants with fixed snow parameters is presented. For each model variant, the CMAES search algorithm and CF6 are applied to optimize soil and runoff related parameters for 1000 grid cells.
The optimization process stopped after 114 (SatBasic) to 417 (SimpleBasic) iterations in each case because the termination tolerance for the cost function value was reached. During the process, parameter values stabilized, e.g. for BergBasic after approximately 40 iterations (A 6).

Tab. 13 lists the default and optimized parameter values of the final calibration for each model variant.
p_et reduces ET in all model variants, except for BudykoGW, where it is increased by 1 %.
For SimpleBasic ***s_fac_simple*** increases the resistance of soil to release water, while simultaneously ***s_exp_simple*** increases the amount of Qs with increasing W. Fig. 29 shows the resulting relation between Qs and W. As can be seen, the resistance of soil is considerably high in the calibrated model, with for instance 179 mm water need to be available to generate 1 mm runoff.

Tab. 13: Default and optimized parameter values applying CMAES and CF6 for 1000 pixels.

Parameter		Default	Simple	Sat		Berg		Budyko	
			Basic	Basic	GW	Basic	GW	Basic	GW
soil	*p_et*	1.00	0.95	0.90	0.91	0.86	0.87	0.88	1.01
	s_fac_simple	0.5	0.0000003	-	-	-	-	-	-
	s_exp_simple	1	2.9	-	-	-	-	-	-
	s_max	300	-	133	166	282	232	163	249
	s_exp_berg	1.10	-	-	-	1.43	2.55	-	-
	s_exp_budyko	0.60	-	-	-	-	-	0.41	1.00
runoff	*q_t*	2.0	3.0	100.0	-	99.7	-	100.0	-
	g_r	0.160	-	-	1.000	-	1.000	-	0.735
	g_d	0.010	-	-	0.019	-	0.025	-	0.009

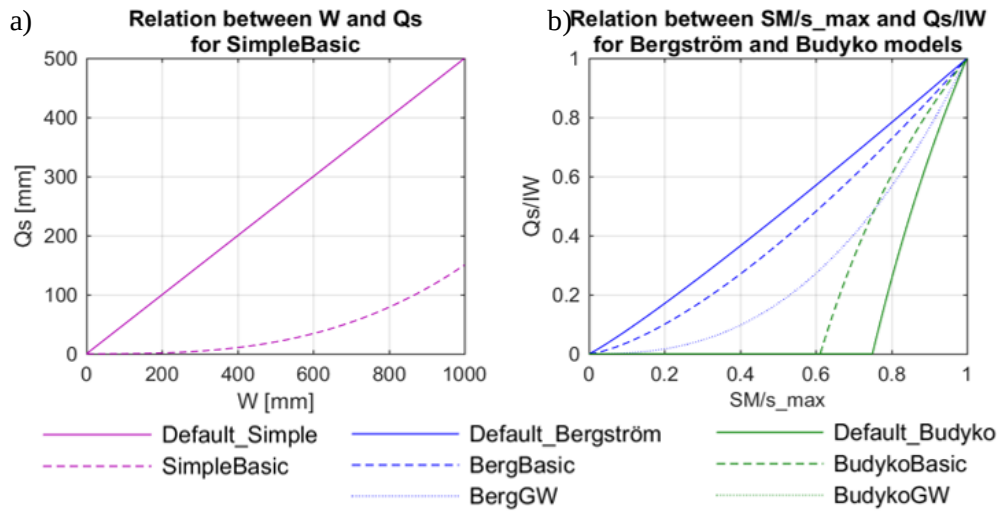


Fig. 29: Influence of the soil parametrization in Simple (pink), Bergström (blue) and Budyko (green) model variants on a) the relation between W and Qs and b) the proportion of IW that contributes to Qs depending on the actual soil moisture deficit SM/s_max. Solid lines represent the default parametrization, dashed lines the calibrated basic runoff variants and dotted lines the calibrated groundwater model variants for Simple models in pink, Bergström models in blue and Budyko models in green. The corresponding parameter values are listed in Tab. 13.

The maximum soil water holding capacity ***s_max*** is decreased for all variants to values between 133 and 282 mm.

In Fig. 29 the influence of the calibrated soil parameters on the proportion of Qs in dependency on the soil moisture deficit is illustrated and contrasted to the default parametrization. For Bergström variants (blue lines) the higher optimized values of ***s_exp_berg*** decrease the amount of Qs if the soil is not saturated (dashed and dotted blue lines). For the Budyko models (green lines), the decrease of ***s_max*** increases Qs under not-saturated conditions compared to the default parametrization, whereby in contrast to Bergström models, no runoff is generated until at least 60 % of the soil water holding capacity is reached. However, BudykoGW in theory shows intermediate behavior between default parametrization and BudykoBasic, yet

discrepancies emerge when analyzing the simulated values. Thus Fig. 29 does not show BudykoGW.

In the basic runoff models, except for SimpleBasic, the recession time of Qs is increased (almost) to its upper bound of 100 days. For SatGW and BergGW, ***g_r*** allocates all Qs to the groundwater storage. The retention time in ground water storage as defined by ***g_d*** varies between 40 days (BergGW) and 111 days (BudykoGW).

5.3 Model Evaluation

In the following section the results obtained with calibrated model variants are presented. First, achieved costs and corresponding ranking of the model variants are reported. Subsequently, performance of Beck-GW models is contrasted to their standard counterpart to examine whether spatial discretization of base flow related characteristics improves model output. Finally, differences in the simulated mean seasonal cycle are shown to evaluate the temporal model performance and areas of high resp. low costs are identified to assess spatial conformity between model and observations.

5.3.1 Model Evaluation and Ranking

Fig. 30 contrasts the costs obtained for calibration (Cal), independent evaluation pixel (Eval) and the whole study area (Global) for each model variant. As SWE costs for the calibration period 2003 – 2010 vary between basic and groundwater runoff variants due to differing initial SWE conditions (A 7), costs depicted here are calculated for the period 2004 – 2010, where SWE costs are identical (Fig. 30c).

For all model variants total, TWS and ET costs for Cal are less than for Eval and Global, whereby global application obtains slightly better results than Eval. Total and ET costs increase by ca. 1% between Cal and Eval, TWS costs by 2-3 %. However, the pattern of model ranking remains the same, except for BudykoGW regarding total costs, which has the best rank for Cal, yet the 3rd rank for Eval and Global.

Regarding global application, BergBasic attains lowest total and TWS costs, followed by BergGW, while highest costs are obtained with SatBasic. Considering ET, BergBasic performs worst, while best values are scored with BudykoGW, which ET costs are 0.05 less than of other model variants (Fig. 30d).

However, similar to model ranking during first investigations (chapter 5.1.3), cost ranges between models are less than 0.1, with 0.06 for total costs, 0.07 for TWS costs, and 0.02 for ET costs excluding BudykoGW.

Costs obtained for global application as well as the corresponding rank are listed in Tab. 14. According to total costs and the weighted average rank, BergBasic is the best performing model variant.

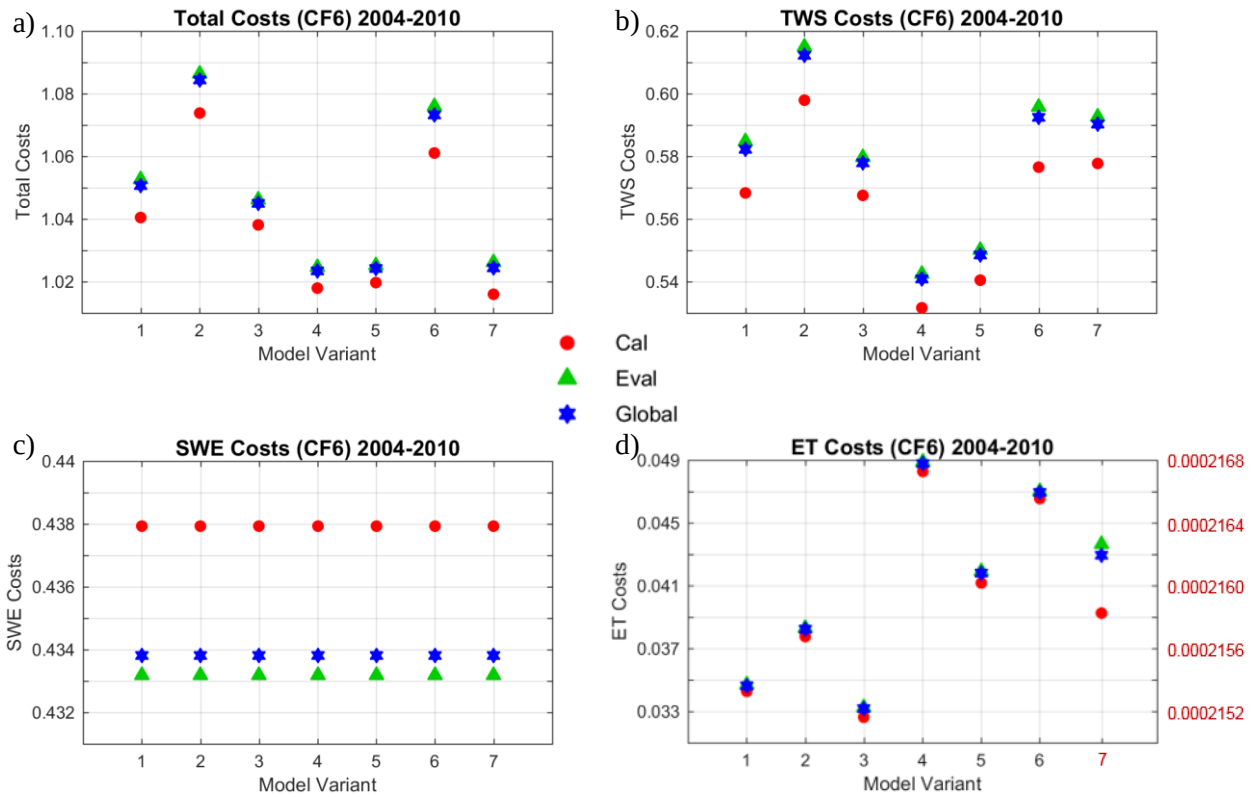


Fig. 30: Comparison of a) total, b) TWS, c) SWE and d) ET costs obtained with optimized model variants for the period 2004 – 2010. (Model variants: 1-SimpleBasic, 2-SatBasic, 3-SatGW, 4-BergBasic, 5-BergGW, 6-BudykoBasic, 7-BudykoGW; optimization for 1000 grid cells using CMAES and CF1).

Tab. 14: Costs for global application of the calibrated models,average ranking according to Gulden et al. (2008) and weighted average ranking (abs represents the absolute value of costs and rk identifies the rank among the model variants respectively). Green color highlights the best value resp. rank; SWE costs for all model variants are 0.434.

	SimpleBasic		**SatBasic**		**SatGW**		**BergBasic**		**BergGW**		**BudykoBasic**		**BudykoGW**	
Costs	abs	*rk*	abs	*rk*	abs	*rk*	abs	*rk*	abs	*rk*	abs	*rk*	abs	*rk*
Total	1.051	*5*	1.084	*7*	1.045	*4*	**1.023**	***1***	1.024	*2*	1.073	*6*	1.024	*3*
ET	0.035	*3*	0.038	*4*	0.033	*2*	0.049	*7*	0.042	*5*	0.047	*6*	**0.000**	***1***
TWS	0.582	*4*	0.612	*7*	0.578	*3*	**0.541**	***1***	0.549	*2*	0.593	*6*	0.590	*5*
mean r		*2.7*		*4.0*		*2.0*		*3.0*		***2.7***		*4.3*		*2.3*
w mean rk		*11.1*		*18.7*		*8.2*		***6.0***		*7.4*		*17.4*		*12.4*

5.3.2 Incorporation of spatial distributed Base flow Information

In the following the costs and parameter values of calibrated Beck-GW models (fixed snow parameters, optimization using CMAES and CF6 for 1000 grid cells) are shown in comparison to their standard counterpart BergGW.

Costs

Fig. 31 displays total costs for BergGW and Beck-GW models obtained for the grid cells used for calibration using default (black) and optimized (red) parameter values. When applying default parameter values, integration of BFI maps improves model performance, while considering K causes higher costs. Also, including K together with BFI2 and BFI4 (Beck2, Beck4) leads to higher costs than the combination of BFI1 resp. BFI3 and K (Beck1, Beck3). After optimization, total costs of Beck-GW models applying solely BFI estimates are similar to the original BergGW model (ca. 1.036), while the performance remains slightly poorer when K estimates are factored in (ca. 1.041).

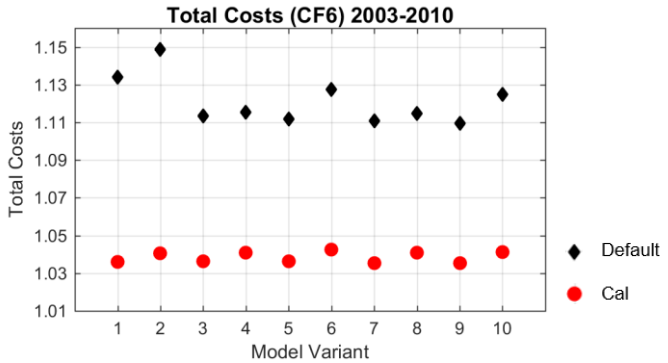


Fig. 31: Total Costs of Beck-GW model variants for 1000 calibration pixel with default parametrization (Default) and optimized parameter values (Cal).Model variants are labeled as explained in Fig. 32.

Optimized Parameter

Fig. 32 contrasts the optimized parameter values of the standard BergGW and the Beck-GW models.

p_et of all variants is approximately 0.87, while ***s_max*** of models including K is around 20 mm higher and ***s_exp_berg*** tends to be lower than with models not applying K. Groundwater recharge rate ***g_r*** is 100 % for both, the standard BergGW and the Beck-GW K model. Similar, ***p_gr*** increases the BFI estimates, while in most cases the parameter reaches its upper bound. The groundwater recession coefficient ***g_d*** of the standard and Beck-GW BFI models is between 0.0245 and 0.0247 d^{-1}, which is equivalent to 41 days. K of all Beck-GW models is decreased to 42 – 44 % of its original value by ***p_gd***, resulting in a mean retention time of 26 to 27.5 days.

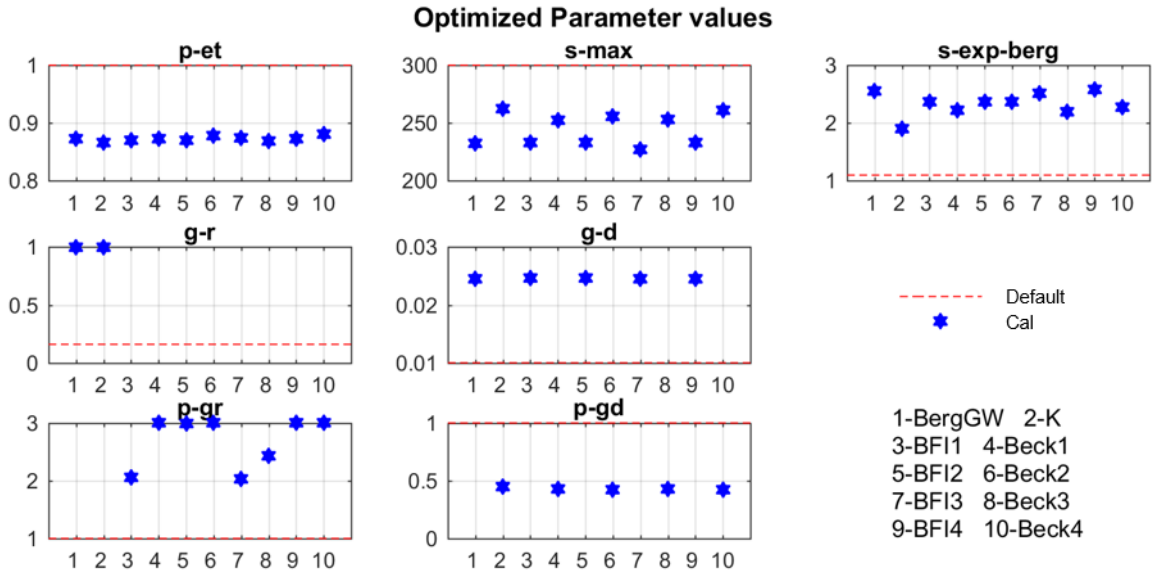


Fig. 32: Optimized parameter values (Cal) of Beck-GW model variants.

5.3.3 Differences in the Mean Seasonal Cycle

To examine whether differences in model structure lead to divergent representation of the temporal dynamic of simulated fluxes and storage components, mean seasonal cycles (MSCs) are investigated. Fig. 33 shows the MSC of ET, Q, SWE as well as storage anomalies in SM, GW and TWS to the average of the period 2004 – 2010 for all model variants and the Beck-GW BFI1 model. As can be seen, the latter is identical to the corresponding standard BergGW variant.

The seasonal dynamic of ET with minimum values during winter and the maximum in June/July is in general captured well by all models, although solely BudykoGW replicates high ET_{obs} values during the summer.

Q is minimal in February/March, increases during spring until reaching its maximum in May/June and decreases again. Except for BergBasic and BergGW that both simulate constant decrease until winter, simulated Q slightly increases again in September and reaches a second, less pronounced maximum in October. Further, apart from SimpleBasic, basic runoff variants obtain higher peak values than their groundwater counterparts, and simulate less Q during winter months. However, it is noticeable that the difference between BudykoBasic and BudykoGW is more pronounced than between SatBasic and SatGW resp. BergBasic and BergGW.

In general, highest evaporation (BudykoGW, SimpleBasic, SatGW) coincidences with less Q during summer months.

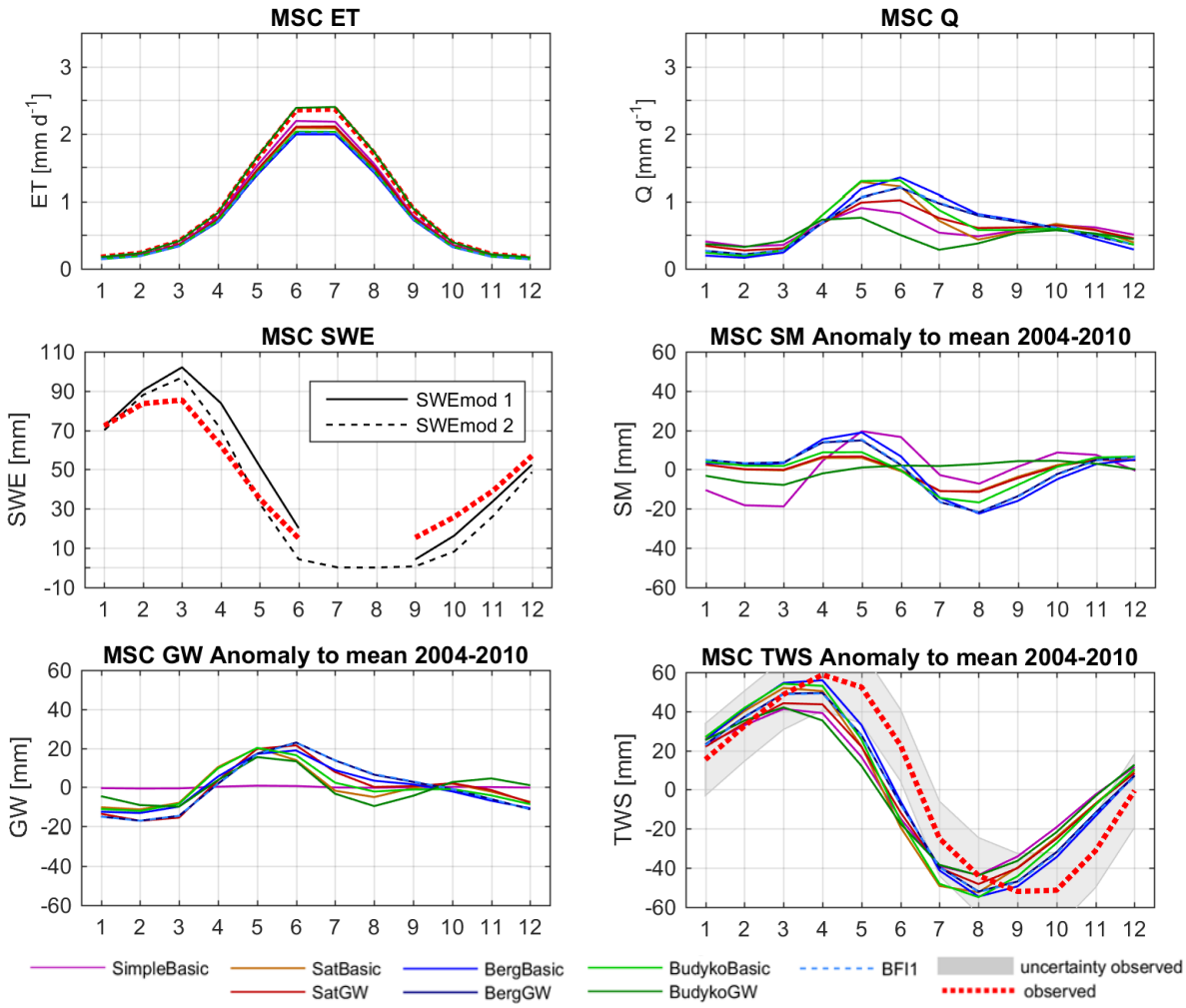


Fig. 33: Mean seasonal cycle (MSC) (2004 – 2010) simulated by the seven standard model variants and Beck-GW-BFI1 model. for ET, Q, SWE, and anomaly of SM, GW and TWS to the corresponding mean of the period 2004 – 2010. SWEmod 1 represents the MSC calculated using data points with available SWE_{obs} while SWEmod 2 illustrates the MSC considering all time steps for all grid cells. The uncertainty estimates of TWSobs (grey shaded area) are provided within GRACE Tellus JPL- RL05M and averaged in the same way as TWS_{obs}.

Due to identical parametrization, all model variants coincidence in the simulated MSC of SWE. They accord with SWE_{obs} in terms of snow accumulation starting in September, reaching its maximum in March and decreasing again until the end of June. Nevertheless, models underestimate SWE during accumulation, and overestimate high snow accumulation as well as snow ablation.

Regarding SM, the absolute values of models differ due to different ***s_max***, yet the anomaly to the mean of the period 2004 – 2010 shows the same pattern for most variants. During winter, SM is around the average value, reaches its maximum in April/May, decreases to the minimum in August before increasing again. This MSC is more pronounced with Bergström variants than with Saturation versions and BudykoBasic. In contrast, SM of SimpleBasic and BudykoGW is minimal during winter, and while SimpleBasic apart from that qualitatively corresponds with the MSC of the other models, BudykoGW shows less variation.

While SimpleBasic shows no variation of GW, the MSC by all other model variants shows a similar pattern. GW is lowest in winter, refills during spring, reaches its maximum in May/June, depletes, except for Bergström variants, until August, increases again slightly resp. stays constant until October, and finally decreases to its minimum in February. Bergström variants differ as they simulate continuous decrease from June to February.

The modelled TWS dynamic with maximum in March/April and minimum in August is similar for all model variants. However, TWS of SatBasic and BudykoBasic decreases faster and the amplitude of SimpleBasic, SatGW and BudykoGW is less than with other models. Compared to TWS_{obs}, all models show shift as TWS_{mod} decreases and increases earlier. Thus, TWS tends to be underestimated in spring and summer and overestimated in autumn and winter. However, TWS_{mod} mostly lies within the range of uncertainty estimates for TWS_{obs} (grey shaded area). This is especially true for Bergström model variants, for which discrepancies are solely in May and June greater than expectable TWS_{obs} uncertainties.

5.3.4 Spatial Distribution of Costs

To assess spatial conformity between model output and observations and to identify areas of good and poor model performance, pixel wise costs (CF6) are calculated for each calibrated model variant and compared.

Between different model variants, the spatial distribution of total costs does not differ largely, with the correlation coefficient being > 0.93. While BergBasic, BergGW and Beck-GW-BFI1 models correspond, SatBasic and BudykoBasic qualitatively show higher costs in Central Siberia, whereas SimpleBasic, SatGW and BudykoGW obtain higher cost values in East Siberia. Regarding ET costs, values increase towards the southern-most regions of the study area (south west North America, Kazakhstan, Mongolia) in all model variants. SWE costs show the same pattern due to identical parametrization. Therefore, spatial variations of total costs among model variants are ascribable to differences in TWS costs.

Fig. 34a shows the spatial distribution of total costs for the period 2004 – 2010 obtained for the BergBasic model. The relative contribution of TWS and SWE to total costs is calculated according Eq. (59) and depicted in Fig. 34b. Hereby positive values indicate stronger influence of TWS costs (red) and negative values of SWE costs (blue), while values around zero (green color) suggest comparable contribution of TWS and SWE costs.

$$relative\ contribution = \frac{TWS\ cost}{Total\ cost} - \frac{SWE\ cost}{Total\ cost} \tag{59}$$

Total costs are highest in the southern areas of the study area, the North American west and north coast, East Siberia, the Central Siberian Plateau and Scandinavia. In general, areas with

low costs are more frequently in Eurasia than in North America. Thereby, costs in North America mostly result from a mismatch of TWS, except for the west coast and northern Canada. Further, TWS costs primarily affect costs in Scandinavia, the West Siberian Plain, North East Siberia and the South East Asian territories of the study area. In Eurasia, SWE costs are the main component of total costs in the East European Plain, Kazakhstan, the Central Siberian Plateau and Eastern Siberia.

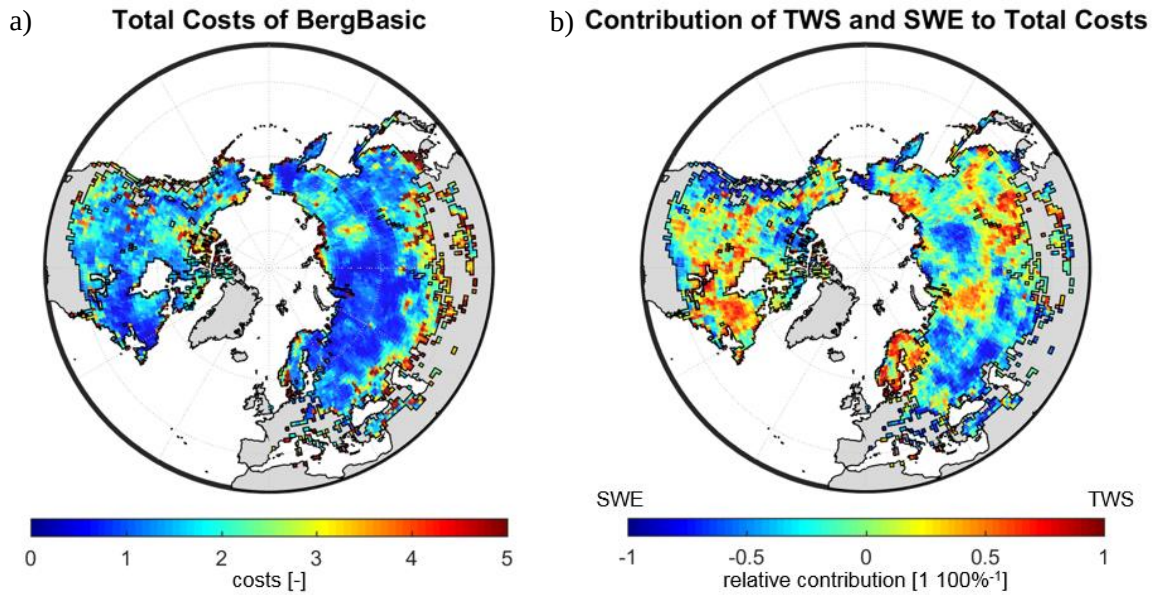


Fig. 34: Total costs of the optimized BergBasic model (a) and relative contribution of TWS and SWE to them (b) for each grid cell for the period 2004 – 2010. In b) negative values (blue color) indicate stronger influence of SWE costs and positive values (red color) indicate stronger influence of TWS costs.

Fig. 35 highlights the location of grid cells with especially poor (red) and good (green) model performance regarding TWS. Accordingly, highest TWS cost tend to occur in coastal and mountainous regions, while lowest costs are mainly obtained in the East European Plain and Central Siberian Plateau.

The percentage of grid cells related to their fraction of water area is depicted in Fig. 36. It shows that 86 % of the study area have less than 5 % water fraction, while 8 % have more than 5 % and another 6 % more than 20 % water coverage. Among the classes of water fraction, the total number of grid cells with high costs is approximately constant. However, as the overall number of grid cells with high water fraction decreases, grid cells with high water fraction are more likely to count among high TWS costs. On contrary, low costs mainly occur for pixels with < 5 % water fraction. Nevertheless, the correlation coefficient between TWS costs (resp. total costs) and the fraction of water area per grid cell is 0.08.

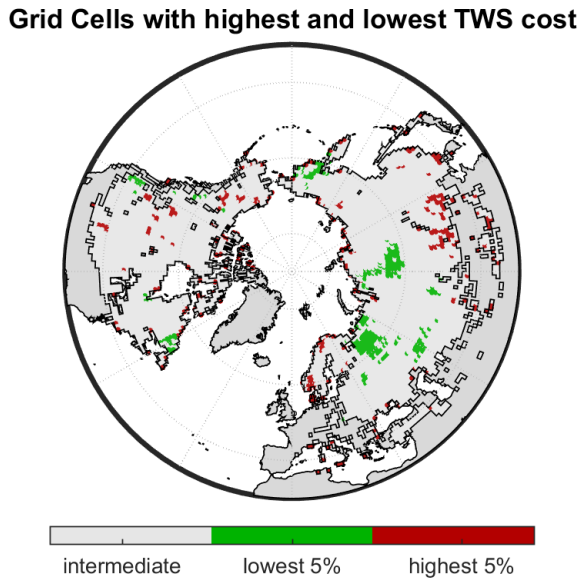


Fig. 35: Spatial distribution of grid cells with highest and lowest TWS costs.

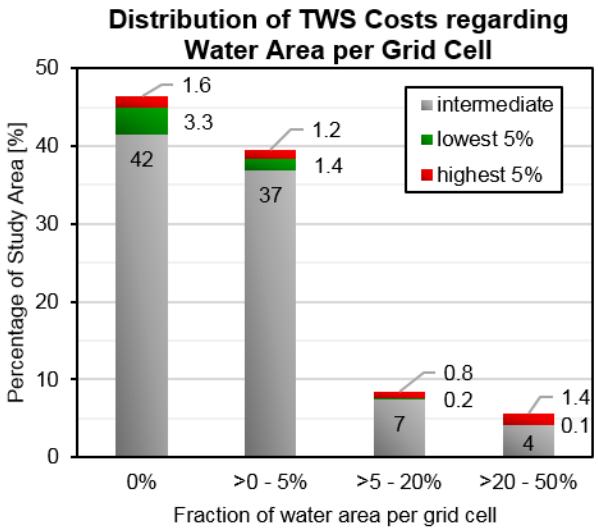


Fig. 36: Fraction of water area per grid cell and relation to TWS costs.

5.4 Application of the best Model Variant

The following section shows the results obtained by application of the BergBasic model variant for the period 2004 – 2010.

First, the MSC of TWS components as well as corresponding anomalies to the MSC are investigated and compared to observations, in order to evaluate whether the model is able to reproduce intra- and inter-annual variability. For both, MSC and anomalies, efficiency criteria are calculated for each grid cell and KG zone to assess spatial conformity. Subsequently, variations in the MSC and anomalies of TWS are attributed to its components to examine spatial differences in their relative contributions on TWS variability. The final section contrasts modelled and observed trends in TWS anomalies, and thus addresses the last research question whether trends in TWS_{obs} can be reproduced.

5.4.1 Mean Seasonal Cycle

The average MSC for the entire study area has been illustrated in chapter 5.3.3. For the MSC of each KG zone see A 12 and A 13.

Thereby Fig. 33 revealed a shift in the average MSC between TWS_{obs} and $TWS_{mod.}$ The spatial distribution of the lag [months] obtained using cross-correlation for each grid cell is depicted in Fig. 37. For approximately 40 % of the study area the MSC of TWS_{obs} and TWS_{mod} are in phase, while around 50 % of all grid cells have a shift of -1 month. A lag of more than 3 months is present in the south east of the Eurasian study area. As it is apparent from Tab. 15, the larger shift coincidences with the arid and boreal semi-arid KG zone.

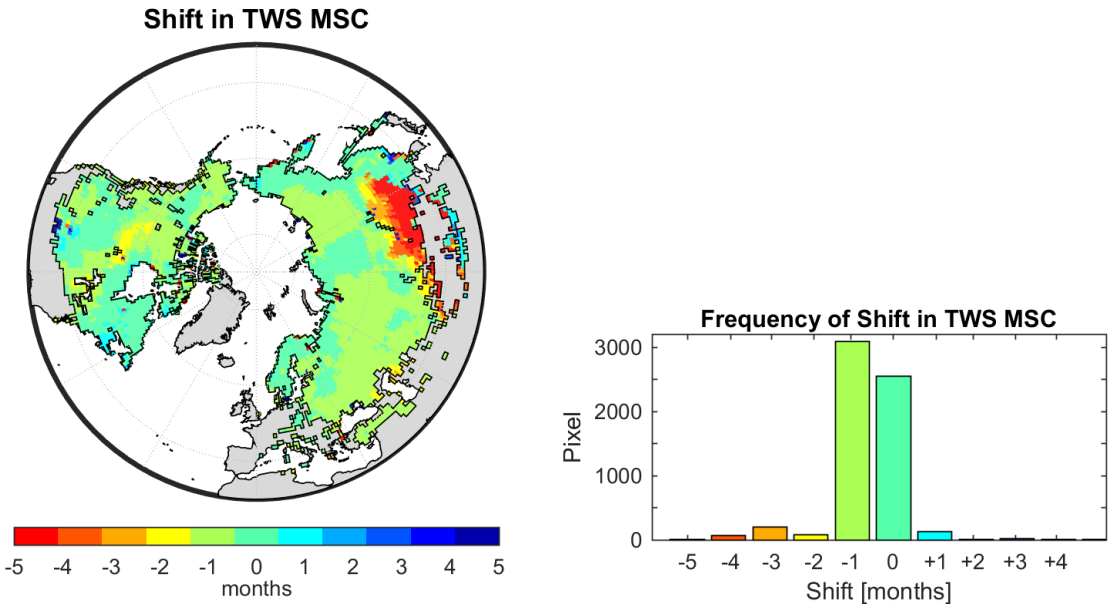


Fig. 37: Spatial distribution and frequency of shift between MSC of TWS_{obs} and TWS_{mod}.

Besides the shift, the average MSC of the entire study area suggests that the seasonal amplitude of TWS_{obs} can be reproduced well (Fig. 33). Simultaneously, the average maximum of the MSC of SWE_{mod} is higher than the maximum of SWE_{obs}. To see whether there are areas where the TWS amplitude is not reproduced and whether the overestimation of mean maximum SWE is imputable to certain areas, Fig. 38 shows the amplitude in MSC of TWS_{obs} and TWS_{mod} for each grid cell, while Fig. 39 maps the average maximum SWE_{obs} and SWE_{mod}.

Regarding the TWS amplitude, Fig. 38 suggest good agreement in the spatial pattern and magnitude for each grid cell. Areas of obvious discrepancy are Scandinavia, where the model overestimates the seasonal amplitude, and the Canadian Archipelago, where the model underestimates the seasonal amplitude.

For maximum seasonal SWE, the spatial pattern qualitatively matches, too, yet the model simulates considerably higher snow amounts in regions where maximum SWE_{obs} > 170 mm (Fig. 39). On contrary, the magnitude for areas of less SWE_{obs} coincidences, except for the North American west coast, where substantial higher snow amounts are simulated.

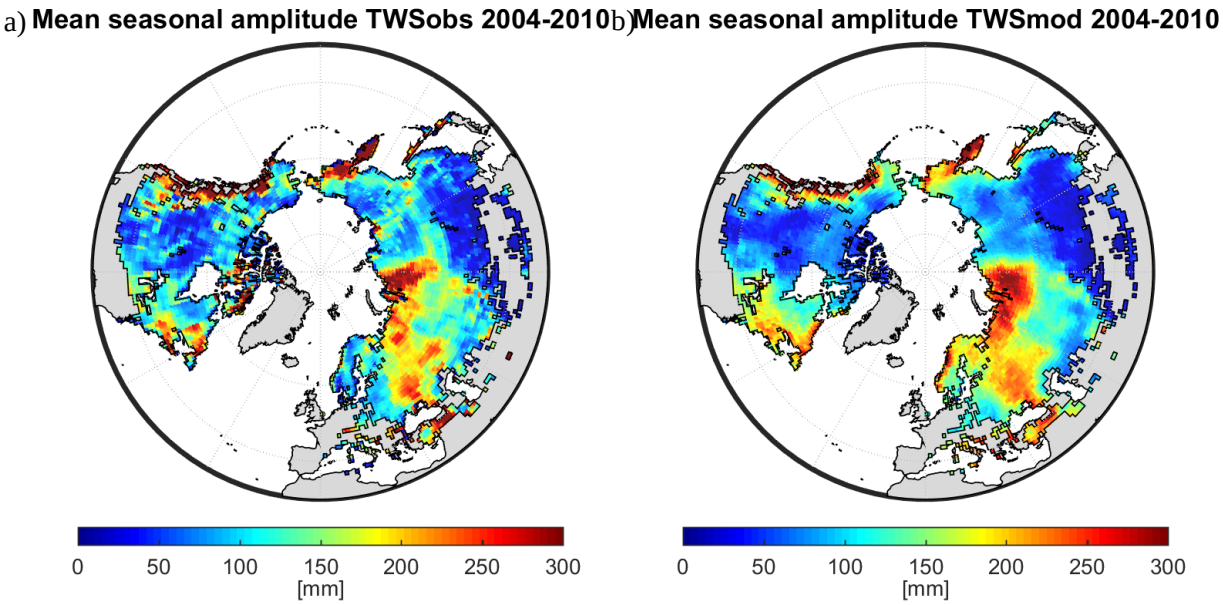


Fig. 38: Mean seasonal amplitude of a) TWS_{obs} [mm] and b) TWS_{mod} [mm]for the period 2004 – 2010.

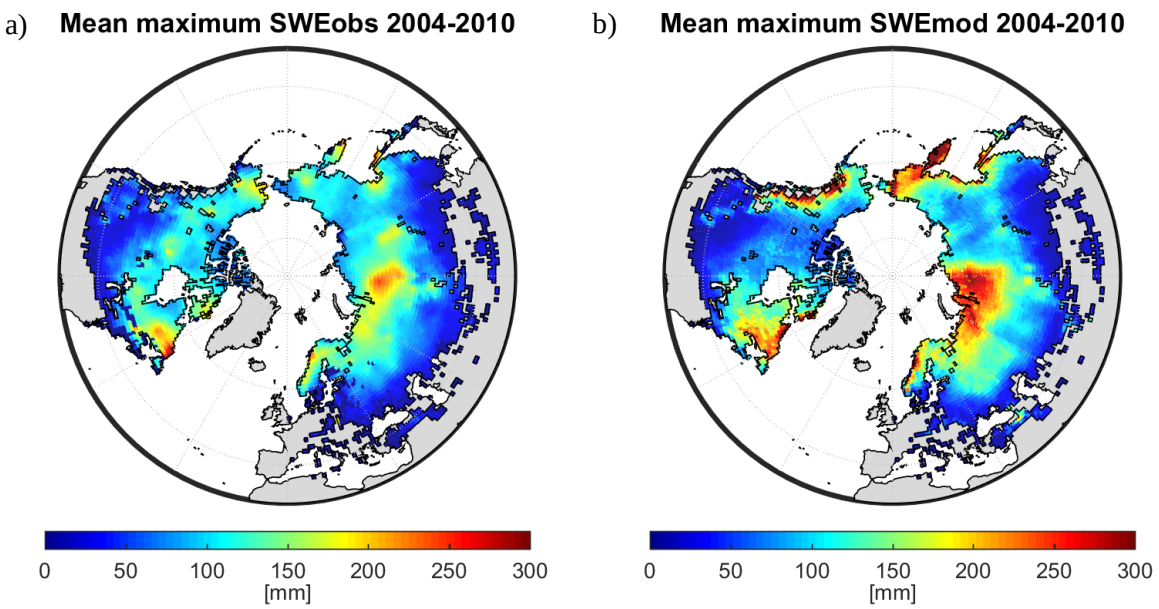


Fig. 39: Mean of maximum annual a) SWE_{obs} [mm] and b) SWE_{mod} [mm] for the period 2004 – 2010.

Tab. 15 lists the correlation, KGE and RMSE of ET, SWE and TWS calculated for each KG zone as well as the whole study area, while Fig. 40 maps the correlation between the MSC of TWS_{obs} and TWS_{mod} resp. SWE_{obs} and $SWE_{obs.}$ For maps of KGE and RMSE criteria see A 9 and A 10.

Regarding TWS, the correlation is with r = 0.93 on average high. Lower correlation is achieved for central North America, the West Siberian Plain, East Siberia, and especially the south east of the Eurasian study area, which is also characterized by the largest shift in TWS MSC. Thereby Tab. 15 attributes the lowest correlation again to the boreal semi-arid and arid zone.

Regarding absolute differences, an average RMSE of 14 mm represents a mismatch of 12 % of the mean seasonal TWS amplitude (110 mm for both TWS_{obs} and TWS_{mod}). RMSE is highest in areas of low correlation and additional along the North American west coast, East Siberia and Kamchatka as well as Scandinavia. This results in relative high RMSE values for the temperate and polar zone (Tab. 15), yet the values remain within the uncertainty range of TWS_{obs}. KGE, as it includes correlation as well as the bias, shows similar patterns as r, yet additional highlights the mismatch in the polar zone and Scandinavia due to absolute differences. However, globally a KGE of 0.93 is achieved.

In comparison to TWS, the correlation of SWE is on average higher (r = 0.95), yet declines towards the south of the study area, with lowest r values again for the arid and boreal semi-arid zone. The global RMSE of 12 mm represents, similar to TWS, 12 % of the mean seasonal SWE amplitude (98 mm). However, RMSE values of 21 mm resp. 17 mm show divergences especially in the polar and boreal zone, whereby highest divergences are obtained for the North American west coast, East Siberia, Kamchatka and the West Siberian Plain. Additional, KGE indicates poorer reproduction of SWE than of TWS, with worst performance in the polar zone.

For ET, r, KGE and RMSE show good agreement between ET_{obs} and ET_{mod}, except for the arid zone. The spatial pattern of r, KGE as well as RMSE correspond to the spatial distribution of ET costs.

Tab. 15: Characteristics of MSC for each KG zone and the entire study area(Global). σTWS_{obs} is the mean uncertainty estimate of GRACE TWS data for each zone, respectively. Criteria for each zone are calculated based on the simulations and observations of each pixel within the zone respectively (not for each pixel and subsequently aggregated).

KG zone	TWS shift [months]	mean σTWS_{obs} [mm]	Correlation r			KGE			RMSE [mm]		
			ET	SWE	TWS	ET	SWE	TWS	ET	SWE	TWS
Arid	-1.5	± 15	0.99	0.92	0.83	0.62	0.64	0.73	0.33	6	13
Temp	-0.6	± 21	1.00	0.98	0.91	0.79	0.64	0.88	0.25	9	20
Bor	-0.5	± 19	1.00	0.98	0.95	0.80	0.82	0.95	0.17	21	14
Bor-sa	-1.9	± 17	1.00	0.90	0.68	0.78	0.72	0.46	0.23	9	14
Polar	-0.4	± 18	1.00	0.93	0.97	0.77	0.50	0.74	0.12	17	15
Global	-0.7	± 19	1.00	0.95	0.93	0.78	0.72	0.93	0.20	12	14

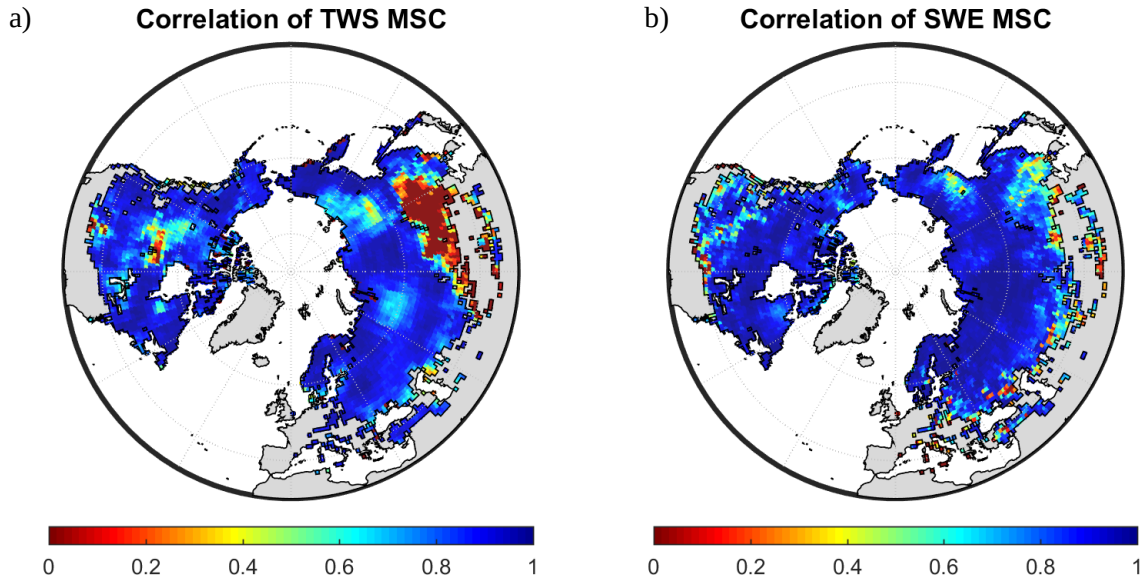


Fig. 40: Correlation of mean seasonal cycle between a) TWS_{obs} and TWS_{mod} and b) SWE_{obs} and SWE_{mod} for each grid cell and the period 2004 – 2010.

5.4.2 Anomalies to the Mean Seasonal Cycle

Similar to MSC, first the time series for the period 2004 – 2010 are investigated to see whether the model can reproduce the temporal dynamic of observed anomalies. Subsequently the spatial pattern is considered to identify areas of good and poor model performance.

The global average anomalies of modelled TWS components (SWE, SM, GW, W, TWS) and corresponding observations are shown in Fig. 41, while the anomalies of each KG zone can be found in A 16 – A 19.

Overall, model performance is poorer compared to the MSC (Tab. 16).

For SWE, the correlation coefficient considering the whole study area is 0.62 (Tab. 16). Especially positive anomalies in the spring 2008 and 2009 as well as in autumn 2010 are not reproduced by the model. This misfit is also shown by RMSE = 26 mm and KGE < 0.

Further, Fig. 41 indicates that variations in mean W are characterized by the same pattern as the anomaly of SM, while GW variations are on average less pronounced.

Regarding TWS, the global r = 0.82 and KGE = 0.48 show better performance than for SWE anomalies (Tab. 16). Thereby general tendencies in fluctuations correspond, while the model does not reproduce short term variations, resulting in an overall RMSE of 69 mm.

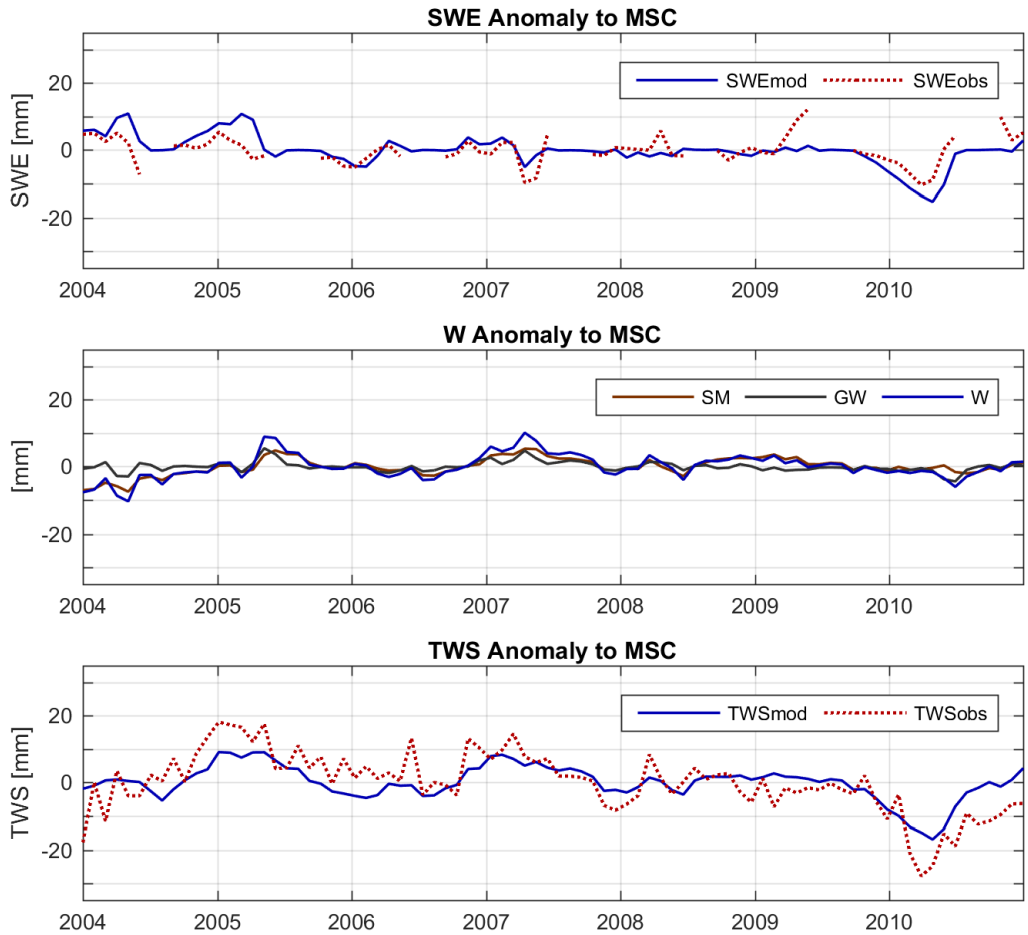


Fig. 41: Average global anomaly to mean seasonal cycle of a) SWE_{obs} and SWE_{mod}, b) SM, GW and W, and c) TWS_{obs} and TWS_{mod}.

Tab. 16: Characteristics of anomaly to MSC for each KG zone and the entire study area(Global). Criteria for each zone are calculated based on the simulations and observations of each pixel within the zone respectively (not for each pixel and subsequently aggregated).

KG zone	Correlation r			KGE			RMSE [mm]		
	ET	SWE	TWS	ET	SWE	TWS	ET	SWE	TWS
Arid	0.93	0.62	0.82	-1.96	-0.39	0.54	0.03	6	7
Temp	0.97	0.77	0.88	-1.79	-0.09	0.66	0.01	5	6
Bor	0.88	0.64	0.83	-0.12	-0.33	0.66	0.00	5	7
Bor-sa	0.67	0.55	0.74	-4.28	-0.12	0.59	0.01	5	10
Polar	0.64	0.41	0.58	-42.77	-2.16	0.33	0.01	9	25
Global	0.93	0.62	0.82	-27.94	-5.81	0.48	0.06	26	69

The spatial distribution of correlation between the anomalies of TWS_{obs} and TWS_{mod} resp. SWE_{obs} and SWE_{obs} is shown in Fig. 42. For maps of corresponding KGE and RMSE criteria see A 14 and A 15.

For both, TWS and SWE, correlation is highest in the central continental regions of the study area and tends to decline towards its borders. Similar, the average correlation of each KG zone reveals low correlation for the polar zone and poor performance in the boreal semi-arid zone.

On contrary to MSC, considering r, anomalies in the arid zone are similar good reproduced as in the temperate and boreal zone. The spatial pattern of KGE coincides with that of r, and KGE as well as RMSE identify highest mismatch in the polar zone. However, both show discrepancies in the arid zone, too.

Concerning ET, the spatial pattern of all efficiency criteria is similar to the one of the MSC and ET costs, yet criteria values are less. Thereby r and KGE indicate lowest agreement for the polar and boreal semi-arid zone, while highest absolute differences are obtained for the arid zone.

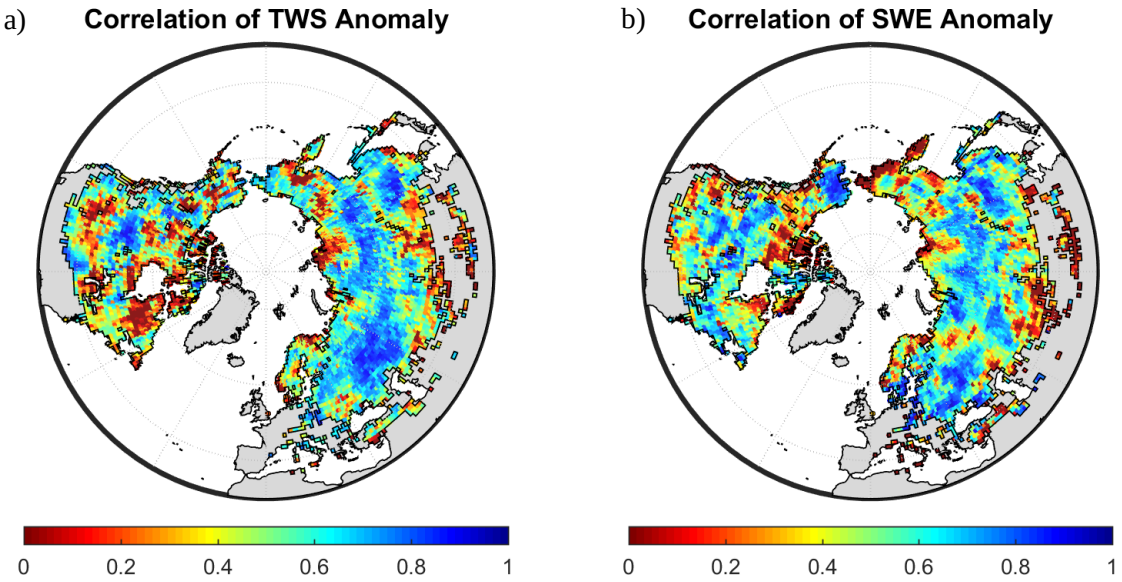


Fig. 42: Correlation of anomaly to mean seasonal cycle between a) TWS_{obs} and TWS_{mod} and b) SWE_{obs} and SWE_{mod} for each grid cell and the period 2004 – 2010.

5.4.3 Composition of TWS

The next section examines the relative contribution of SWE, SM, GW and W to TWS variability. In order to see whether the influences change on different time scales, the intra-annual variability (MSC) as well as the inter-annual variability (anomalies to MSC during the period 2004 – 2010) are considered.

Mean Seasonal Cycle

Fig. 43a plots the modelled variance against the observed variance in the MSC of TWS. As can be seen, the magnitude of seasonal variance in general can be captured by the model, yet it does not reproduce highest variability.

Fig. 44 shows the spatial distribution of variance in the mean seasonal cycle of TWS_{obs} and TWS_{mod}. Visual comparison and a Spearman's rank correlation $\rho = 0.71$ indicate good reproduction of the spatial pattern. Both show highest variance along the North American west coast, for North East Siberia including Kamchatka, the West Siberian Plain, the East European

Plain, Anatolia, and North East North America. Observed high variance on the Canadian Arctic Archipelago is not reproduced by the model, while it simulates high variance for Scandinavia which is not measured by TWS_{obs}.

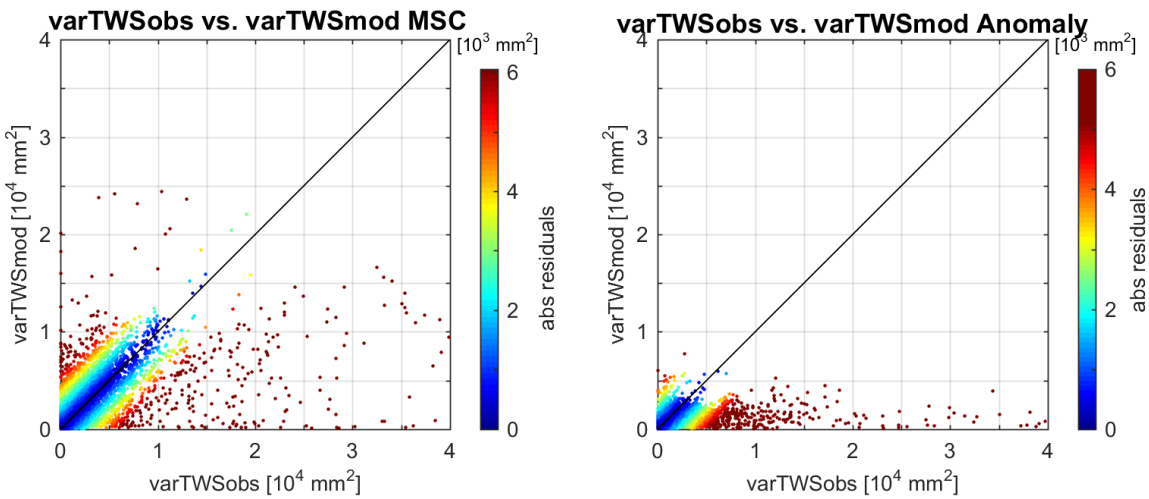


Fig. 43: Pixel wise variance of TWS_{obs} vs. TWS_{mod} in a) the mean seasonal cycle, and b) anomalies to the mean seasonal cycle. Dark red color marks absolute residuals that exceed the 95th percentile.

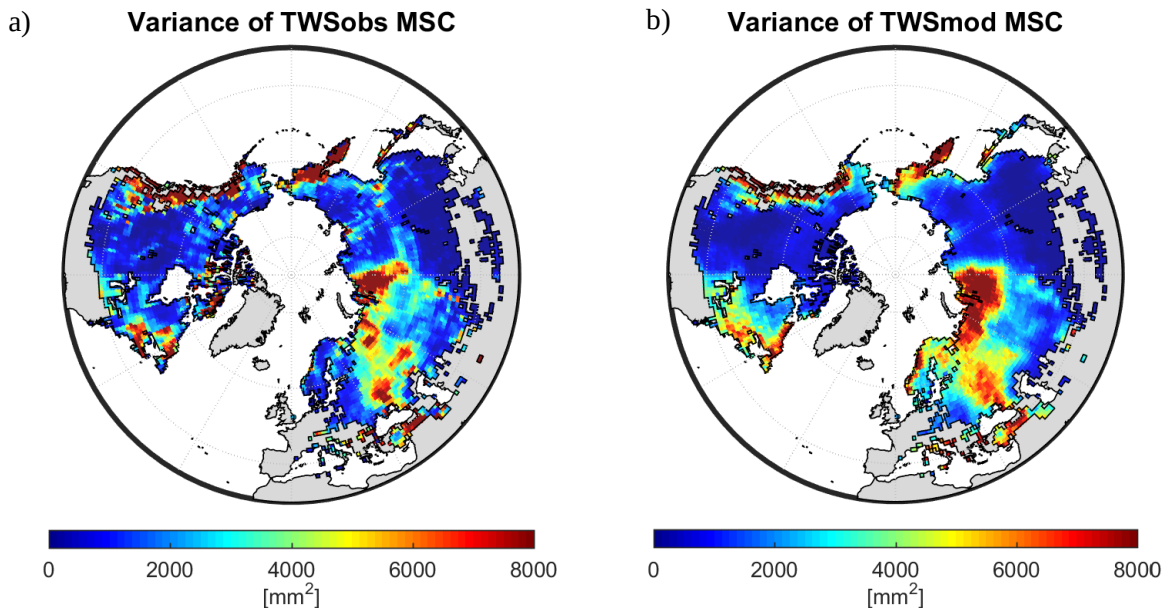


Fig. 44: Variance in the mean seasonal cycle of a) TWS_{obs} and b) TWS_{mod} for each grid cell and the period 2004 – 2010.

If two times the covariance between W and SWE is low, meaning the variances varW and varSWE do not correlate, variability in TWS can be seen as the combined effect of varW and varSWE. Indeed, the covariance between W and SWE in general is low, although A 21 shows a slightly negative correlation between SWE_{mod} and W in areas of highest SWE_{mod}.

However, Fig. 45a shows the difference between varW/varTWS and varSWE/varTWS, so that negative values (blue color) indicate a greater influence of variability in SWE on intra-annual TWS variations, while positive values (red color) identify areas where TWS variability is mainly caused by variability in SM and GW (which sum up to W).

Thus, Fig. 45a suggests that variations in the MSC of TWS in northern regions are mainly influenced by variations in SWE, while south of approximately 50° latitude variations of W prevail. Compared to North America, the influence of W in western Eurasia reaches more northwards and the transition to SWE dominated regions is more gradual.

As W is composed of SM and GW, its variability can be explained in the same way. Thus, Fig. 45b shows the relative contributions of varSM and varGW on varW, with negative values (blue color) indicating a greater influence of varGW and positive values (red color) indicating dominance of varSM. Accordingly, Fig. 45b suggests that variability in W is primarily caused by variations of SM. The influence of simulated GW solely prevails on the Canadian Arctic Archipelago, the northern coast of Eurasia, East Siberia and East China.

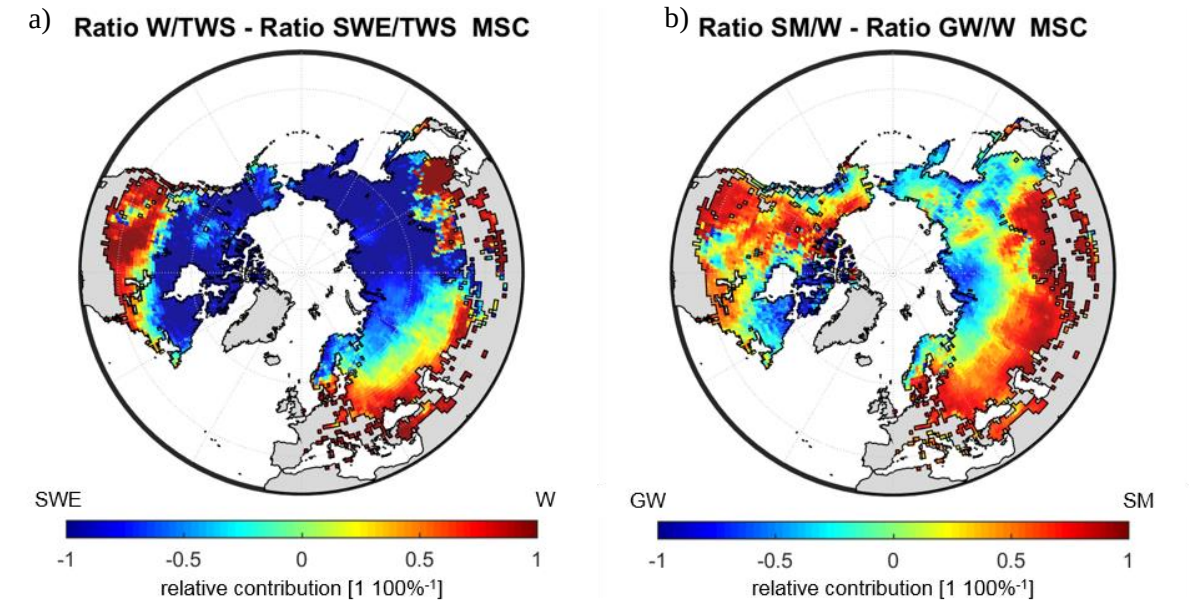


Fig. 45: Difference between a) varW and varSWE normalized with varTWS and b) varSM and varGW normalized with varW of the MSC.

Anomaly to the Mean Seasonal Cycle

Fig. 43b plots the modelled variance against the observed variance in the anomaly of TWS. Compared to the seasonal variation (Fig. 43a), the simulated inter-annual variability of TWS is in general less. Thereby, the model considerably underestimates high TWS variance.

The spatial distribution of the variance in TWS_{obs} and TWS_{mod} anomalies is shown in Fig. 46. Qualitatively the spatial pattern can be roughly reproduced ($\rho = 0.36$). Similar to MSC, areas

of high variance are the North American west coast, Kamchatka, the West Siberian Plain, the East European Plain, Anatolia, and North East North America including the Canadian Arctic Archipelago, while low variance is measured in south eastern regions of the study area. In contrast to MSC, high variability is present in central North America, and the distribution in general appears patchier.

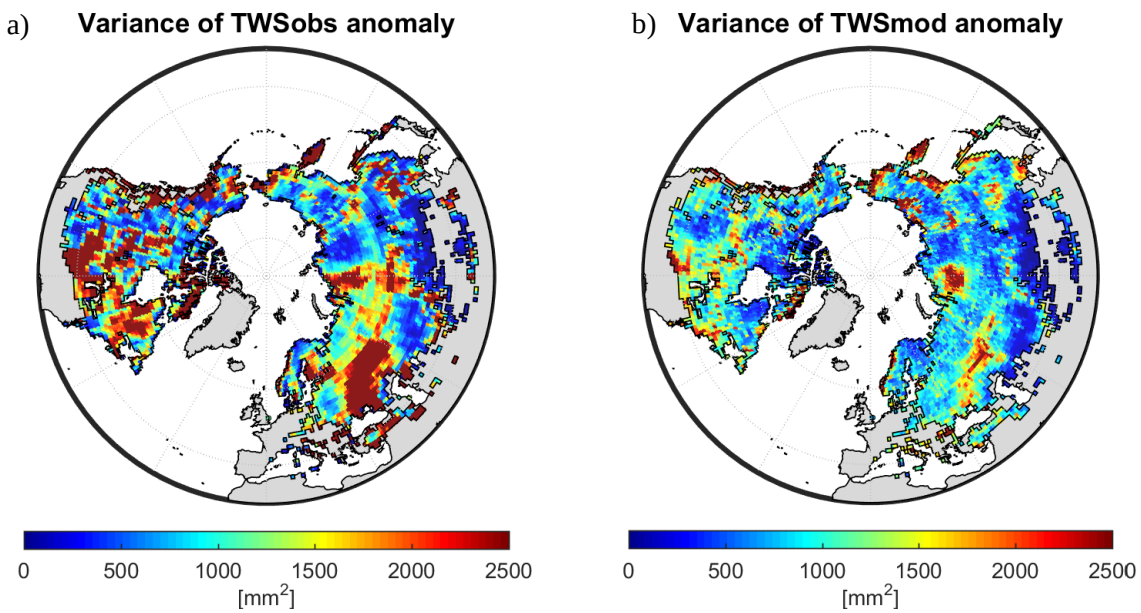


Fig. 46: Variance in the anomaly to the mean seasonal cycle of a) TWS_{obs} and b) TWS_{mod} for each grid cell and the period 2004 – 2010.

Similar to MSC the covariance between W and SWE resp. SM and GW is mostly low (A 23), and thus variability in TWS resp. W can be attributed to its components, respectively.

Fig. 47a shows the difference between varW/varTWS and varSWE/varTWS, so that negative values (blue color) indicate a greater influence of variability in SWE on inter-annual TWS variations, while positive values (red color) identify areas where TWS variability is mainly caused by variability in W. As Fig. 47a indicates, inter-annual variability in TWS is largely dominated by W, while influence of SWE variations is only present in on the Canadian Arctic Archipelago, the northern west coast of North America, North East Siberia and the northern West Siberian Plain.

Additional, Fig. 47b displays the relative contributions of varSM and varGW on varW, with negative values (blue color) indicating a greater influence of varGW and positive values (red color) indicating dominance of varSM. As can be seen, the variability of W in areas where it dominates TWS variation is predominated by SM, while in regions where TWS is mostly influenced by SWE, W varies due to variations in GW.

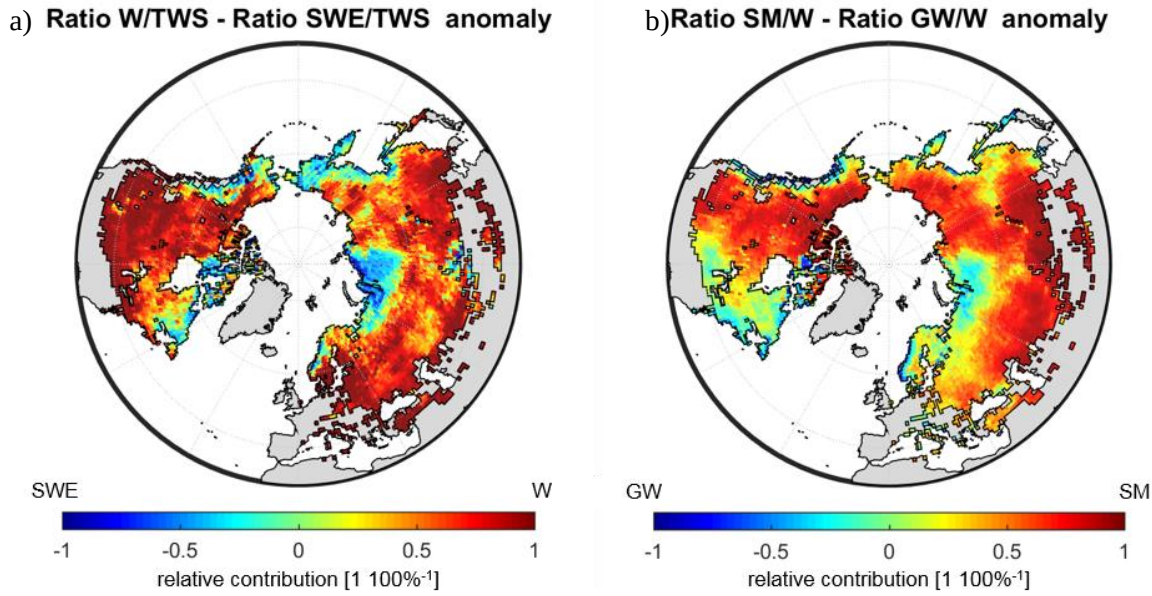


Fig. 47: Difference between a) varW and varSWE normalized with varTWS and b) varSM and varGW normalized with varW of the anomalies to MSC.

5.4.4 Trends in TWS

To examine trends in the observed and simulated TWS, trends obtained by the Mann-Kendall test and applying Sen's slope for each grid cell are qualitatively classified in not significant ($p > 0.05$), positive (Sen-slope > 0) and negative (Sen-slope < 0). Fig. 48 shows the spatial distribution of these classes for TWS_{obs}, while Fig. 50 highlights areas of correspondence and deviations between observation and model regarding trend direction as well as significance. The corresponding confusion matrix lists the number of grid cells per class for each classification and the resulting accuracy. According to Fig. 48, negative trends in TWS_{obs} are present for the West Eurasian Steppe, East Siberia, Alaska and large parts of central Canada including the Canadian Arctic Archipelago. Positive trends on contrary are observed for central North America, the Labrador Peninsula, as well as for large parts of Eurasia.

Overall, these trend directions can qualitatively be reproduced with an accuracy of 42 %. Thereby 49 % of the significant negative trend and 52 % of the significant positive trend are captured. Thus, the model obtains better agreement with observations than when trend classes are assigned randomly, which on average produces an overall accuracy of 33 % (A 27). While the spatial pattern of agreement/disagreement is heterogeneous (Fig. 50), it shows that the large-scale negative trends in the West Eurasian Steppe, East Siberia, Alaska and Canadian Arctic Archipelago, as well as less widespread positive trends in East Europa, and central Asia and North America can be simulated.

However, Fig. 49 contrasts the obtained Sen-slopes for TWS_{obs} and TWS_{mod} anomalies. Similar to the TWS variance in anomalies, the model tends to underestimate both, positive and negative trends.

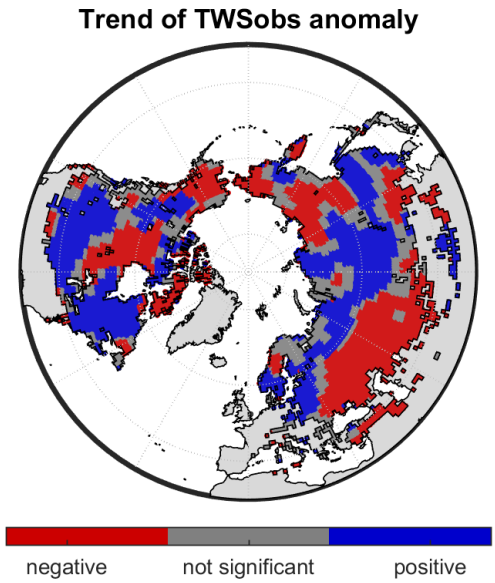


Fig. 48: Qualitative trends in the anomaly to the mean seasonal cycle of a) TWS_{obs} and b) TWS_{mod} for the period 2004 – 2010.

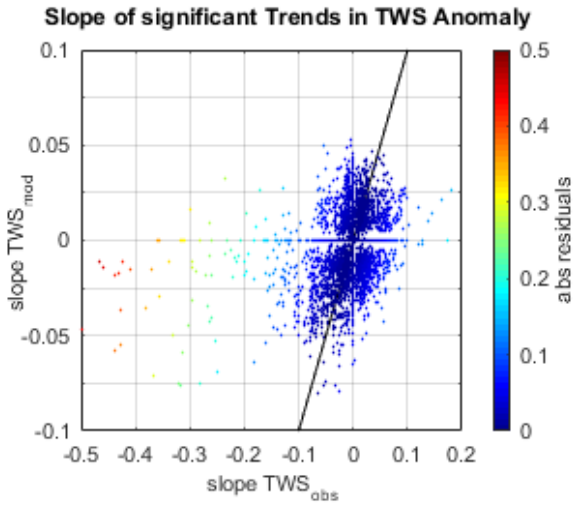
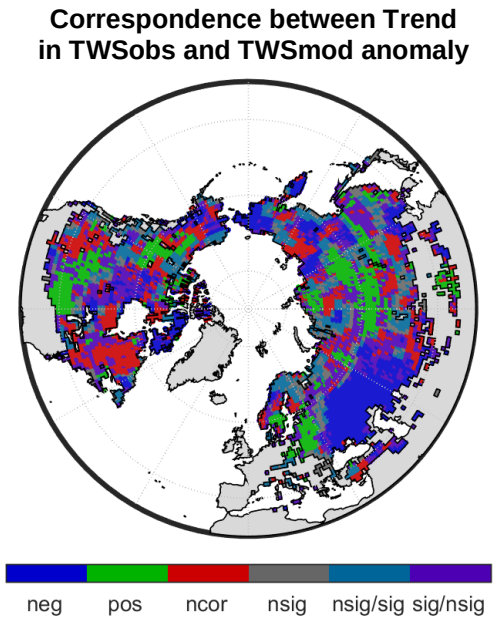


Fig. 49: Sen-slope of significant trends in TWS_{obs} vs. TWS_{mod} anomalies. Black solid line: 1:1 line.



		obs			Total	User Acc
		neg	not sig	pos		
mod	neg	1147	673	471	2291	0.50
	not sig	552	553	382	1487	0.37
	pos	634	870	908	2412	0.38
Total		2333	2096	1761	6190	
Producer Acc		0.49	0.26	0.52		**0.42**

Fig. 50: Correspondence between trends in TWS_{obs} and TWS_{mod} anomalies to the mean seasonal cycle (left) and confusion matrix (right).neg/pos – correspondence in significant negative/positive trends; ncor – no correspondence in significant trends; nsig – correspondence in not significant trends; nsig/sig – not significant trend in TWS_{obs} but significant trend in TWS_{mod}; sig/nsig – significant trend in TWS_{obs} but not significant trend in TWS_{mod}.

6 Discussion

The following chapter discusses the results presented in the previous section and aims to answer the research questions of this thesis.

First, some general sources of uncertainty and errors are outlined that are basic for further interpretation. The subsequent section addresses the second research issue "which optimization method is most suitable". Therefor the tested optimization methods are reviewed, the most suitable is identified, and the quality of the final calibration is discussed. Ensuing, the calibrated model variants are compared regarding their parameter values, their simulated mean seasonal cycles, as well as their overall performance, in order to answer research question one "which model formulation is 'best'". The forth part of this chapter finally focusses on the scientific research questions by analyzing the spatial and temporal model performance in comparison to GRACE TWS and GlobSnow SWE products, and discussing the composition of TWS variations. Eventually, some remarks for future work are proposed.

6.1 Sources of Errors and Uncertainties

Since every model is an approximation of the real system, uncertainties are inherent in modelling results (Fischer, 2013). They originate from multiple sources, including a general lack of knowledge and process understanding, the applied forcing and calibration data, model structure and parameters, as well as the quality of the model calibration (Moriasi et al., 2007). Thereby it should be noted, that while subject to its own uncertainties, calibration to a certain degree compensates errors and inconsistencies of the input data and structural problems originating from simplified process representation (Müller-Schmied et al., 2014).

Input Data

Regarding the input data, global estimates are derived from in-situ measurements, remote-sensing based observations and/or reanalysis, each of which are subject to specific uncertainties that are transferred when linking them to combined products (Lettenmaier et al., 2015). Thus, uncertainty in global data sets arises from a sparse observational network, measurement errors and gaps due to missing data as well as errors and assumptions in retrieval algorithms and methods (Döll et al., 2015). Furthermore, inconsistencies between the used data products may introduce additional errors in the model results (Sood and Smakhtin, 2015).

Among the meteorological forcing data of hydrological models, uncertainty is highest for precipitation products (Döll et al., 2003). This study uses the GPCP-1DD product, which combines precipitation estimates from different satellites and employs a gauge analysis to control for the bias between satellite retrievals and ground measurements. While currently multi-satellite data "provide the best source of precipitation information" (Lettenmaier et al.,

2015), issues remain due to orographic effects, heterogeneous surfaces especially in coastal zones, difficulties in distinction between precipitation and cold land as well as for quantitative estimates of snow fall because of diverse radiative effects depending on the shape of snowflakes (Lettenmaier et al., 2015). Further, although bias-adjusted, small scale and high intensity precipitation events may not be captured (Behrangi et al., 2011), and uncertainty remains for large regions of the study area due to a sparse observational network. On the other hand, the observational data used for correction itself introduces errors, as gauge data is characterized by a systematic under-catch because of losses due to wind, wetting, evaporation, infiltration as well as blowing and thrifting of snow. Frequent light rainfall, windy conditions and errors during snow fall lead especially in arctic regions to substantial underestimation of precipitation (Rudolf and Rubel, 2005).

Similar to the precipitation data, the applied evapotranspiration product is subject to errors of its components, as the eddy covariance measurements and the MODIS products, as well as of the retrieval method itself. Thereby it is known for poor estimation of inter-annual variability (Jung et al., in preparation).
Since existing global climate data sets differ in their data basis, processing and resolution, simulations are highly sensitive to the used meteorological forcing (Guo et al., 2006). Similar, Müller-Schmied et al. (2014) found that seasonal TWS variations simulated with WaterGAP 2.0 are mostly influenced by climate forcing data, rather than by model structure or auxiliary physio-geographic information. To determine the influence of the data streams applied in this study, further analysis may compare model results when forced with other data sets.

Besides meteorological forcing, also the data used for calibration and evaluation are subject to specific errors. While their uncertainty estimates have been included in the optimization process, errors need to be considered when interpreting the obtained modelling results as well. GRACE TWS estimates suffer from measurements errors in the satellite-instruments and orbit determination, errors due to incomplete reduction of non-hydrological mass variations (e.g. post-glacial rebound, seismic deformations) during processing and leakage errors from signals of adjacent areas (Güntner, 2008; Landerer and Swenson, 2012). The latter are inherent to the product, as GRACE measurements are continuously and thus a filter function, or as in the version used in this study, mascons have to be applied to extract mass variations for a certain area. Thereby, processing seeks the optimal compromise between noise reduction (accuracy) and signal loss (spatial resolution). Scaling the GRACE data with the provided gain factors, allows to account for the filtering effect on the signal and interpret it at sub-mascon resolution (Watkins et al., 2015). Since appropriate ground truth data on TWS is missing, the uncertainty of the final TWS product can only be estimated based on the measurement errors and simulation

studies that account for the leakage error. In general, errors increase from high to low latitudes, as well as with decreasing similarity to the hydrological signal in the surroundings (Güntner, 2008).

As highlighted before, accurate satellite-based SWE measurements remain a problem due to saturation effects of the signal, and especially in areas of heterogeneous surface and snow distribution as mountains and forests (Lettenmaier et al., 2015). Kim et al. (2009) found passive microwave retrievals of SWE most accurate in North American and Eurasian tundra and prairies, that are characterized by relatively thin, cold snow packs and little topographic variability or dense vegetation. While GlobSnow SWE as an assimilation product shows clear improvement in the accuracy compared to pure satellite estimates, underestimation for deep snow conditions (e.g. at the end of winter) and overestimation for SWE < 30 mm (e.g. when accumulation starts resp. at the end of melt season) has to be considered (Hancock et al., 2013; Liu et al., 2014).

Model Structure

Uncertainties in the model structure are related to the applied model algorithms and the neglection of relevant processes. In general, the model error is supposed to decrease with increasing model complexity, while simultaneously the associated increasing data demand increases errors caused by input data (Müller-Schmied et al., 2014).

As this study aims to develop the simplest feasible model, numerous hydrological processes are simplified or ignored. Among others, all model variants do not consider coincident occurrence of rain- and snowfall, rainfall interception on snow, liquid water capacity of snow, interception, surface overland flow when infiltration capacity is exceeded, capillary rise and other surface-ground water interactions, as well as surface water storages as lakes and rivers, human water use and lateral water flow from one grid cell to another. Partly, neglection is reasonable considering the scale and intended spatial resolution. For instance, according to Hagemann and Dümenil (1997) capillary rise is insignificant on the global scale, and although human water use causes strongest variations in the mean annual TWS in arid and semi-arid regions (Döll et al., 2015), Müller-Schmied et al. (2014) found its incorporation less decisive compared to model forcing and other parameters.

Among the included processes, predefined threshold temperatures for snow fall and melt may introduce problems, as both are solely based on literature and a range for transition of rainfall to snowfall resp. of increasing melt rate would better reflect reality. Further, all model variants include temporal retardation of runoff, either by explicitly considering a groundwater storage and distinguishing between direct runoff and base flow (groundwater variants), or by introducing a delay component and implicitly assigning restrained water to groundwater storage

(basic variants). In both cases, this one retardation component faces and thus has to comprise various (intermediate) storages and delay times of the real system.

Since this study uses an ET product, it does not distinguish between transpiration and evaporation and assumes that evaporation from the canopy and surface water as well as the effect of land cover resp. land use is included in the ET product.

Additionally, sub-grid spatial heterogeneity and, more general, spatial diversity of physio-geographic characteristics as soil properties, land cover and altitude that influence hydrologic responses significantly are neglected. The effect of introducing such spatial distributed information on model performance is investigated with the Beck-GW models (see chapter 6.3.4).

Model Parameters

Additional to model structure, uncertainties originate from the interdependencies between different model parameters. To assess this uncertainty due to model parametrization, to identify the effect of individual parameters on the modelling result and to determine the 'sensitive' parameters, it is advisable to conduct a sensitivity analysis, which includes multiple model runs with diverging parameter sets (Fischer, 2013). This has not been done in this study, but should be included in further research, as in particular parameters of the snow component are highly interconnected (see chapter 6.2.2). Therefore, also the correlation matrix between the parameters has to be inspected.

Further, all parameters are optimized for a global uniform value, and thus do not reflect spatial differences, as would be for example expected for the soil parameters due to spatial variability of soil properties.

Quality of Calibration

Uncertainties in the calibration arise from the applied optimization method itself, meaning the suitability of the search algorithm and the cost function, as well as from the optimization settings (e.g. termination criteria). Further, the choice of parameters to be optimized and their predefined parameter bounds influence the quality of the calibration result. As several optimization methods have been tested, these issues will be discussed in the following section.

6.2 Model Optimization

The subsequent paragraphs review the choice and remaining issues of the applied optimization method as well as the (final) model calibration. Thus, they address the second methodological research question.

6.2.1 Optimization Methods

In general, different ways of multi-criteria calibration exist. This study combined all criteria in one cost function and performed one-objective calibration, in order to reduce computational time, resulting in one 'optimal' solution. As the contribution of the individual criteria to total costs reflects their conceptual relevance – ET costs < SWE costs < TWS costs – summing them up equally is acceptable and no additional weighting necessary.

Out of the tested search algorithm, the local solvers were faster, yet obtained higher costs compared to CMAES, which suggests that they converged in a local minimum. While GSFmincon needed the longest computational time due to multiple starting points, it did not considerably improve costs. Moreover, the differences between the algorithms' solutions are small, yet optimized parameter values differ largely. This indicates a highly complex parameter space with several local optima, which makes a global search algorithm unavoidable. From the applied global solvers, CMAES needs less computational time, obtains lowest costs and in general is supposed to perform superior for heterogeneous parameter spaces (Hansen, 2014). Therefore, it is chosen as the most suitable search algorithm.

However, investigation of the trade-offs between the model's ability to fit all criteria would yield further insights in model behavior and parameter uncertainty due to input data and model structure. Other multi-criteria approaches as the Non-dominating Sorting Genetic Algorithm II, that for instance has been used by Werth et al. (2009) and Xie et al. (2012), provide a Pareto-front with 'equally good' parameter combinations and thus allow such examination (Vrugt et al., 2003a; Xie et al., 2012).

Regarding the tested cost functions, it was shown that ET_{obs} needs to be included as a constraint. Otherwise, the calibrated models tend to neglect ET as surplus water just as well can drain, since runoff is not limited. Additional constraining the model by discharge data would force simulation of ET in order to match TWS and Q, so that ET could probably be excluded from the cost function.

Further, trimming of the data was found to be necessary to diminish the effects of extreme values, which likely are outliers resp. result from measurement errors. Another common approach in hydrologic modelling is to use logarithmic transformed efficiency criteria, that put less emphasis on high values (Krause et al., 2005). This makes sense when the aim is to evaluate discharge, and in particular low flood conditions. However, TWS represents anomalies to a time-mean baseline and thus in general high (negative and positive) values need to be reproduced in order to simulate the amplitude correctly. Regarding snow dynamics, logarithmic variants would be suitable in terms of considering the underestimation of deep snow conditions, yet they simultaneously put too much weight on overestimated low SWE_{obs} values. Thus,

applying an efficiency criteria weighted by the individual uncertainty of the constraint is most convenient. As using wMSE resulted in unrealistic simulations, wMEF is found to be most appropriate to evaluate the model performance. Among the tested ways to account for the uncertainty in $SWE_{obs,}$ CF6 finally is able to assign costs in the intended way. Although it does not explicitly include the overestimation of thin snow layers, the non-varying absolute value of σSWE reflects higher uncertainty for low values of SWE_{obs}. However, it should be noted that the saturation of the SWE signal among others depends on the snow grain size and the stratification of the snow pack (Takala et al., 2011), and thus setting a fixed threshold to determine when SWE_{obs} does not provide reliable estimates is somewhat arbitrary.
Nevertheless, the findings show that the search algorithm CMAES and cost function CF6 based on the wMEF criteria is the most suitable optimization method.

6.2.2 Model Calibration

During the first investigation of model results when testing different cost functions, it was found that the parameter values substantially vary between the model variants, especially for the snow component. This variation of snow parameters in particular reflects well-known problems of parameter equifinality, because all model variants use the same snow component and no structural differences exist. It is very likely that the snow component is over-parametrized, as four out of the five snow parameters regulate the reduction of SWE – either by scaling the amount of snow fall (***p_sf***), simulating sublimation (***sn_alpha***), or modelling snow melt in dependency on temperature (***m_t***) or radiation (***m_r***). Further ***sn_alpha***, ***m_t***, and ***m_r*** are related to temperature resp. net radiation, which in turn correlate. Thus, SWE_{mod} can be reduced by various parameter combinations to fit with SWE_{obs}, which is proven by comparable SWE costs of the different model variants. This increases uncertainty in assigning the relevance of individual processes. So, although a spatial and temporal distinction between sublimation and snow melt would be desirable, the current structure of the snow component does not allow reliable interpretations. Therefore, further studies should either use the traditional day-degree approach instead of the restricted one, or refrain from the attempt to explicitly consider sublimation. Alternatively, considering discharge in multi-criteria calibration may decrease uncertainty in snow parametrization as it would allow better control on the amount of SWE that supplies input water in form of snow melt and the proportion that leaves the system through sublimation. In either case, investigation of the parameter's correlation and interdependencies should be conducted.
In order to reduce parameter equifinality and to facilitate the focus on the effects of differing soil and runoff parametrization, it has been decided to fix snow parameters during final model calibration. Therefor the parameter set of the best performing model during first investigations, namely BergGW, has been chosen. Although this parameter set reflects the majority resp.

average of the other model's snow parameter, the predefinition introduces uncertainty, as for instance the parameter set is calibrated for 100 grid cells instead of 1000 as in final calibration, and on the other hand other model variants may perform better with individual snow parameters. Nevertheless, detailed investigation of snow related processes is beyond the scope of this thesis, as the focus lies on runoff generation and dynamics of TWS – for which reliable representation of the snow pack as it is achieved with this parameter set is sufficient.

Regarding overall model performance in terms of costs, final calibration could improve conformity with TWS_{obs} and SWE_{obs}. Thereby higher costs for evaluation grid cells and global application are common. However, the difference is minor and thus suggests that parameters optimized for 1000 grid cells are representative for the whole study area and models are not over-calibrated.

Final model calibration terminated after a few hundred iterations because the tolerance for the cost function values was reached, which can also indicate a very flat region in the parameter space (Xu, 2002). However, as the parameter values stabilized after solely 40 iterations, this may also relate to the predefined snow parameters, that don't allow much variability in the soil and runoff parametrization to gain an optimal fit with TWS_{obs}. Consequently, predefined snow parameters enable direct comparison of the soil and runoff parametrization, but to exploit the model's full potential all sensitive parameters need to be calibrated.

Further, subsequent investigations should use identical initial storages, in particular regarding the snow conditions, for basic and groundwater variants. A shorter initialization period is reasonable, but the last year(s) should coincide to allow consistent comparison.

6.3 Comparison of Model Variants

To identify the model formulation that is most appropriate to simulate TWS dynamics, seven model variants have been implemented, calibrated and compared as part of this study. The following section aims to answer

1) how the model variants differ after final calibration in terms of their parameters,
2) their mean seasonal dynamics and
3) their overall performance, as well as
4) whether spatial varying of base flow related parameters facilitates improvements in model performance, and finally
5) which model variant is 'best'.

6.3.1 Calibrated Model Parameters

As the snow parameter set is predefined, the individual model variants do not differ regarding their simulated snow dynamics. For all variants solid precipitation is reduced by ***p_sf***, although

an increase would be more likely in view of the typical undercatch especially of snowfall. However, reduction on contrary accounts for sublimation of frequent light snowfall and high winds, that are common in high latitudes (Derksen et al., 2014). Sublimation itself is not considered using the chosen snow parameter set, as the sublimation factor ***sn_alpha*** is set to zero. Thus, no reduction of SWE is simulated as long as temperatures remain below 0 °C. Instead, the snow pack is only reduced by melt, whereby both, ***m_t*** and ***m_r*** (almost) reach their upper bound, theoretically leading to fast decrease of SWE_{mod} as soon as temperatures exceed 0 °C. High values of both parameters indicate that on global average both, temperature and net radiation, determine reduction of the snow pack. The rapid decrease introduced by ***m_t*** and ***m_r*** is smoothed by a relatively high value of ***sn_c***, because melt is scaled down if SWE_{mod} is below the 207 mm threshold that is associated with complete snow coverage of the grid cell. As a result, melt is highest under deep snow conditions while a shallow snow pack vanishes comparatively slow.

Among the final calibrated models, the parametrization of SimpleBasic generates runoff solely if the amount of available water is relatively high. In contrast to the other model variants, the retardation ***q_t*** of 3 days is short, so that the comparatively low water release leaves the system nearly immediately. While BergBasic almost shows a linear relation between soil moisture deficit and the runoff generating proportion of IW, in BudykoBasic at least 60 % of the maximum water holding capacity has to be reached to generate runoff. BergGW is intermediate between both approaches. ***s_max*** of SatBasic and SatGW are comparatively low, which indicates the necessity of runoff generation even when few water is available. In the basic versions, runoff is delayed up to 100 days, which equals to the upper limit and thus it is likely that delay time would be increased if it was not constrained. This assumption is countered by the fact that although the calibrated SatGW and BergGW completely allocate Qs into the groundwater storage, the delay time within this storage is short. For BudykoGW the retardation time is comparable to the basic versions, but it simulates Qdir, which compensates the differences in total Q to the other model variants.

According to Hagemann and Dümenil (1997), overland flow with delay times of several days to few weeks is dominant among flow processes in regions that are highly affected by snow melt. Thus, the calibrated models in reflect the main runoff process resp. water retardation associated with the study area. However, base flow, that accounts for constant discharge during winter and in dry areas, usually is denoted by lag times from few months to several years in hydrologic models with 0.5° spatial resolution (Hagemann and Dümenil, 1997). This long-term retardation thus is clearly not included in the present models.

Besides, detailed investigation of the model results showed, that the parametrization of BudykoGW causes unexpected model behavior. After calibration, the parameter ***s_exp_budyko*** reaches its upper bound of 1, so that the divisor in one exponent in the infiltration computation

(Eq. (29)) becomes zero. For division through zero the MATLAB environment generates infinity, which eventually results in calculation of infiltration as the difference between ***s_max*** and SM. Consequently, this bug needs to be fixed by adjusting the upper parameter bound of ***s_exp_budyko*** in order to allow appropriate conclusions about the performance of the intended BudykoGW approach.

6.3.2 Differences in the Mean Seasonal Cycle

In general, all model variants simulate comparable mean seasonal dynamics of water fluxes and storages, in particular for ET and TWS anomalies, as they were used to constrain the model, and for SWE because of identical parametrization. Besides, minor differences exist regarding timing and quantity, which will be discussed in the following.

Regarding snow dynamics, the models coincidence with the MSC of SWE_{obs}. Thereby underestimation at the beginning of accumulation as well as overestimation of deep snow conditions are reasonable when considering the uncertainty of the SWE_{obs} estimates, which overestimate thin snow layers and saturate for deep snow packs. The obtained overestimation during ablation from March to June on contrary suggest that melt is not modelled fast enough. As the melt factors ***m_t*** and ***m_r*** are high, this primarily relates to ***sn_c***, that reduces the decrease of SWE_{mod} for medium and thin snow layers. Nevertheless, the simulated SWE dynamics are within the assumed uncertainty of SWE_{obs}.

Monthly ET is highest in June and July due to high net radiation resp. temperature, highest water supply from snow melt and summertime precipitation, as well as due high plant productivity. In the model variants, ***p_et*** reduces especially these high ET_{obs} values, while solely BudykoGW with ***p_et*** = 1.01 does not substantially modify ET_{obs}.

Although ET is not decreased, SM of BudykoGW shows almost no seasonal variations, as its near water holding capacity throughout the year. This missing seasonal cycle can be attributed to the bug in the soil parametrization. SM, as simulated by the other model variants, is maximum in April/May due to input from snow melt, minimum in August due to high ET, and then increases again as consequence of summertime precipitation and decreasing ET. These seasonal variations are more pronounced in the Bergström variants because they allow reduction of SM in terms of Qs for greater soil moisture deficits. During winter, SM is on its average value, as it receives no resp. less input due to prevailing solid precipitation and is simultaneously not reduced by ET or Qs – in the saturation approach because the maximum water holding capacity is not reached and in Bergström and Budyko variants because of missing liquid input. As Qs in the SimpleBasic approach is not dependent on liquid input but on water availability, negative anomalies are simulated in winter as the water storage still releases runoff. Moreover, for SimpleBasic SM represents all available water, including soil moisture, ground water etc. The implicitly derived GW on contrary can be understood as delayed runoff, e.g. in

channels. Therefore, the shorter retardation time of 2 days makes sense, because long-term retardation is already included in the calculation of Qs. Accordingly, SimpleBasic has no MSC of 'groundwater', which remains constant throughout the year. The other model variants on contrary show the depletion of the GW storage in winter, which is equivalent to the negative SM anomaly of SimpleBasic. Apart from depletion in winter, GW anomalies show the same seasonal pattern as SM. Thus all model variants confirm in the seasonal cycle of W, which is characterized by negative anomalies in winter due to drainage and positive in spring as result of infiltrating snow melt (Niu et al., 2007).

Similar, runoff is minimal in winter due to less liquid input as determined by low temperatures. It increases in spring due to snow melt and reaches its maximum in May/June when nearly the whole snow pack is melted. This represents a delay of 2 – 3 months in comparison to maximum snow cover, and corresponds to the retardation time defined by ***q_t*** and ***g_d***. Like SM, Q shows a second maximum in October that is related to precipitation maxima in July/August. In the Bergström variants summertime runoff is comparatively high and thus the second maxima less pronounced, because runoff is already generated for greater soil moisture deficits. In general, summertime Q is less for models with higher ***p_et*** values, as ET reduces SM and thus less water is available for runoff generation. Further, higher peak values in the basic runoff variants are caused by the logarithmic form of the retardation equation – a value of 100 days for ***q_t*** means that within 100 days all water leaves the system as runoff, but the amount is higher during the first days. In contrast, groundwater runoff variants apply a linear storage that constantly releases base flow, resulting in higher wintertime Q. The noticeable disagreement between simulated Q of BudykoGW and BudykoBasic can be inferred to the changed algorithm by ***s_exp_budyko*** = 1.

Due to snow accumulation in February and March and subsequent melting, TWS in mid- and high-latitudes of the Northern Hemisphere is maximum in April. This positive signal remains until June, followed by negative values due to runoff and ET, before mass starts to accumulate again in October (Rangelova et al., 2007). In contrast, all models show a lag of approximately one month, so that TWS_{mod} decreases too rapidly, reaches its minimum in August and then starts to accumulate too early. This suggests insufficient representation of retardation (Güntner, 2008), but model structure and parametrization cannot be the only reason as a longer delay time theoretically is possible in the groundwater variants. As this feature is common to all model variants, it will further be discussed within analysis of the model application in chapter 6.4.1. In addition to the phase, some models show discrepancies in the amplitude. Nevertheless, the simulated TWS anomalies are mostly within the uncertainty range of TWS_{obs}, and with BergBasic being closest to TWS_{obs}.

6.3.3 Differences in Model Performance

Regarding their overall performance in terms of costs, the ranges between individual model variants are small, which suggests, that all are equally suitable to simulate SWE and TWS dynamics. Comparatively low ET costs achieved for the BudykoGW variant result from the maintenance of ET values, as ET_{obs} is not reduced by ***p_et***.

In spatial terms, similar patterns of costs are obtained for all model variants, indicating that all approaches are equally suited to simulate TWS dynamics for the study area and good resp. poor performance in one region is not primary influenced by differences in the soil and runoff algorithm. This is further supported by the remaining ranking of model variants for calibration grid cells, evaluation grid cells and global application. An exception is the degradation of BudykoGW, which is presumably related to the bug in its algorithm,

Alltogether, the calibrated Bergström approaches, which were among the best performing variants throughout the testing of optimization methods, obtained the lowest costs after final calibration. Thereby BergBasic outperformes BergGW in terms of costs, as well as in terms of computational time needed for calibration.

6.3.4 Incorporation of spatial distributed Base flow Information

To examine whether spatial distributed information improves model performance in comparison to global uniform parametrization, global estimates of base flow index BFI and base flow recession constant K obtained by Beck et al. (2015) have been incorporated in the BergGW model. These products are supposed to provide useful information for large-scale groundwater recharge estimation and parametrization, and indeed, Beck-GW models applying BFI (Beck-BFI1-4) estimates for determination of groundwater recharge showed superior performance compared to their standard counterpart with default parametrization. However, if K estimates are integrated (Beck-K, Beck1-4), results decline. BFI and K as defined by Beck et al. (2015) consider all slow runoff components, such as snow and ice melt, interflow and base flow. Integrating BFI into the model to determine the percolating resp. delayed portion of Q is feasible, because the groundwater storage in BergGW has to compensate all lagged runoff. Similar, K comprises various delay times and among others those associated with snow. Since the snow storage is explicitly simulated within the model, this causes poorer performance when K estimates are applied.

In order to account for these discrepancies and to reduce inconsistencies between the input data, BFI and K are allowed to be adjusted by the scaling parameters ***p_gr*** resp. ***p_gd*** during final calibration. Calibration improved the performance of all model variants, that afterwards achieve similar costs. Thereby the calibrated Beck-BFI1-4 models perform equally good as the calibrated standard BergGW, while model variants using K still obtain slightly worse results.

Regarding the soil component, the calibrated models have similar values of ***p_et***, ***s_max*** and ***s_exp_berg***. For Beck-BFI1-4 the scaling parameter ***p_gr*** increases the fraction of percolating water, for BFI2 and BFI4 as much as the parameters' upper bounds allows, and less for BFI1 and BFI3 that already have higher values than the other BFI estimates. As consequence, similar to the standard BergGW, all Qs percolates, resulting in similar costs. Thus, calibration diminishes the effects of integrating spatial distributed information.

K is characterized by retention times of approximately 14 days in high- and mid-latitudes and 3 days in southern, arid regions (A 2). As the calibrated parameter ***p_gd*** decreases the recession coefficient to 42 – 44 %, the delay time for the majority of the study area is increased to 34 days and thus approximates the delay of standard models, while the lag in arid regions is only 8 days. Although a spatial varying delay time is physically reasonable considering diverse pedologic and geologic properties, it results in poorer model performance in terms of costs, probably because of shorter retention especially in arid regions.

In conclusion, no improvement regarding simulation of TWS by applying spatial distributed characteristics related to the percolation rate or recession coefficient is achieved, as in the calibration process the additional information is diminished. These data may provide useful information to partition runoff or constrain HMs to discharge, but do not aid in the context of this study.

6.3.5 Résumé

In conclusion, all model variants have comparable costs, in terms of overall performance as well as in terms of spatial distribution. Further, they simulate comparable mean seasonal dynamics of water fluxes and storages, whereby minor differences are related to the individual parametrization. Thereby, incorporation of spatial distributed base flow parametrization yields no additional advantage. Thus, all model variants are in general equivalently well suited to simulate TWS. Despite strong resemblance, the models applying the Bergström approach achieved the best agreement with GRACE TWS. Although an explicit groundwater component as in BergGW enables simulation of longer water retardation including base flow induced runoff during winter, BergBasic showed superior performance regarding TWS. As BergBasic additionally requires calibration of one parameter less and thus less computational time, it can be considered as the most appropriate model formulation in terms of consistency with GRACE TWS and therefore has been applied for detailed analysis of TWS dynamics in this thesis.

6.4 Model Application

The following section discusses the results obtained by applying the BergBasic model and aims to answer the scientific research questions of

1) how well the model can reproduce temporal and spatial patterns in the observations,
2) how the TWS is composed and how this composition varies spatially and temporally,
3) as well as whether the model is able to reproduce trends in the GRACE TWS signal.

6.4.1 Spatial and Temporal Model Performance

In general, the BergBasic performs fairly well for the majority of the study area.
Regarding overall model performance, East Siberia and Scandinavia are identified as regions of high errors, that are mainly caused by differences between TWS_{mod} and TWS_{obs}. Additional, the model performs poor at the Central Siberian Plateau and along the North American west and north coast. The poor performance in these regions, as well as in other areas with relatively high total costs, is mainly due to a disagreement in SWE dynamics.
Besides, model performance decreases towards the South of the study area. On the one hand, this is due to a greater mismatch between ET_{obs} and ET_{mod}, on the other hand the model seems not able to reproduce low TWS variations in areas that receive few annual snow.
Higher errors regarding ET in the arid and boreal semi-arid zone can be attributed to less ET_{mod} in summertime. ET_{mod} is reduced by the scaling factor ***p_et***, which, since it is globally optimized, is adjusted to the needs of the wetter and cooler boreal and temperate zones that capture larger areas. Compared to other zones, the precipitation and thus soil moisture that provides water for ET_{mod} is less in the arid and boreal semi-arid environments and thus leads to this greater misfit.

To further assess the models ability to reproduce spatial and temporal patterns of the GRACE TWS and GlobSnow SWE observations, the mean seasonal cycle and inter-annual anomalies are investigated. Besides the overall performance, regions of discrepancies are highlighted and possible reasons for these differences discussed.

Mean Seasonal Cycle

Regarding snow dynamics, the model in general overestimates annual maximum SWE, especially in regions where SWE_{obs} is high, as the Central Siberian Plateau, and what accordingly are the regions of highest RMSE. However, considering the underestimation of deep snow conditions by SWE_{obs}, this overestimation by the model is reasonable. Solely along the North American west coast moderate SWE_{obs} values are highly overestimated, which probably relates to underestimation of TWS_{obs} variations (see below). Additionally, the KGE highlights the polar zone as a region of considerable discrepancies. This zone in general suffers from issues in the GlobSnow retrieval algorithm due to the influence of (permanent) ice layers, ice/snow mixed pixels, and mountainous terrain (region in Asian mountains classified as polar).

Regarding the temporal correlation SWE is reproduced better than TWS, among others because the seasonal cycle is in comparison less influenced by uncertain data streams as summertime precipitation and ET (while snow fall is scaled and thus adjusted to better fit SWE_{obs}).
Nevertheless, mean seasonal variations of TWS_{obs} are in general well captured by the model, as proven by high correlation and KGE values as well as by RMSEs within the uncertainty range of TWS_{obs}.
On average, BergBasic reproduces the seasonal TWS amplitude. However, the amplitude is underestimated in arid and semi-arid regions, as well as for the Canadian Archipelago, and overestimated for Scandinavia. Underestimation in the arid and boreal semi-arid zones simultaneously results in highest RMSE values for these regions. Possible reasons that may result in underestimation of the seasonal amplitude are the scaling of high summer time ET to less amounts by ***p_et***, underestimation of precipitation input (Xie et al., 2012), neglected input from adjacent wetter areas and mountain ranges, as well as neglection of fluctuations in deep groundwater and human water extraction. Additionally, according to Hagemann and Dümenil (1997), base flow with retardation times from few months to several years plays an important role in the hydrological cycle in dry regions with rare precipitation, yet the applied BergBasic model is not able to model such long delay times.
Underestimation in the area of the Canadian Archipelago can relate to problems in satellite measurements of light snow fall (Lettenmaier et al., 2015) and thin snow packs containing hoar layers (Derksen et al., 2014), and thus result from underestimation of both, the precipitation forcing as well as the SWE_{obs} constraint. Additionally, underestimation of TWS_{obs} variations may be largely influenced by leakage errors in the GRACE signal that includes variations of adjacent (sea) ice masses (Rangelova et al., 2007). Similar, overestimation in Scandinavia, which is also associated with high RMSE values, is mainly due to leakage errors, that reduce the GRACE estimate due to smaller signals over oceans (Niu et al., 2007).
High RMSE is also obtained for the North American west coast and Kamchatka, which analogous can be inferred to leakage errors, but additionally coincidences with highest RMSE for SWE. Thereby the model overestimates SWE_{obs} and underestimates TWS_{obs} variations, which indicates that it is only capable to fit TWS variations by simulating high variations in SWE. Using both observations as constraints results in intermediate estimates and errors regarding both, SWE_{obs} and TWS_{obs}. Thus the discrepancies mainly result from data inconsistencies in these regions. Thereby errors originating from the SWE_{obs} retrieval algorithm are likely, as these are mountainous regions with diverse topography, forests, lakes and rivers and additionally under the influence of the coast. As both data sets are subject to retrieval uncertainties, it is difficult to ascertain which is more reliable. However, high annual precipitation suggests that SWE_{obs} is more uncertain, or, less likely, that the simple model

structure lacks processes other than snow accumulation and melt to generate high TWS variability.

Besides divergences in the amplitude, differences in the phase of the MSC are more obvious. Such a systematic delay between GRACE TWS and hydrological models has already been obtained by earlier studies, where, depending on the region, modelled TWS preceded 1 – 6 weeks (Rangelova et al., 2007; Schmidt et al., 2008; Werth et al., 2009). Commonly, these discrepancies are attributed to missing water storage compartments and associated delay of runoff, as well as to shortcomings in the meteorological forcing and/or the GRACE data itself (Güntner, 2008).

As the differing phase is inherent to all tested model variants, model structure in terms of insufficient water retardation cannot be the primary reason. Moreover, a less negative anomaly of TWS from September to February coincidences with underestimation of SWE_{obs} during snow accumulation, while underestimation of positive TWS_{obs} anomalies in spring correlates with overestimation of SWE_{obs} during ablation from March to June. Since optimization intends to fit both variables, TWS_{mod}, which in wintertime mainly consists of SWE_{mod}, in its monthly amount is intermediate between TWS_{obs} and SWE_{obs} in order to achieve the best performance compared to both variables. This again suggests, to some degree, inconsistencies between GlobSnow and GRACE data streams. Considering the uncertainties of the observational data, SWE_{obs}, and thus to some extent still SWE_{mod}, is overestimated for the beginning of accumulation, leading to less negative TWS_{mod} than observed. Further, the underestimation of deep snow conditions by SWE_{obs} causes an underestimation of positive TWS_{obs} anomaly. During spring and until August, the negative anomaly in TWS_{mod} is more pronounced than in TWS_{obs}. Thus, the retardation of melt water (and precipitation) seems to be insufficient. In the groundwater variants a longer delay time would have been possible, which further could compensate other missing (surface) storages by implicitly including them in their parametrization. However, groundwater variants did not show superior performance, which suggests that regional differing aspects of retardation with diverse delay times exist, that are not coverable by one linear storage alone. For example, in the northern and polar regions retardation over frozen soil may play a role. Further, routing and inundation along large rivers would delay water export, but vary spatially in dependence on the drainage network (Güntner, 2008). For instance, Alkama et al. (2010) showed that variations in river storage highly effect TWS variations, especially in cold regions where river water is partly frozen during winter.

Besides these discrepancies in spring, high precipitations in late summer and simultaneously decreasing ET cause increasing simulated mass accumulation starting from August, while TWS_{obs} reaches its minima not until September. Despite higher uncertainty of TWS_{obs} in these

months, summertime precipitation may be overestimated or ET underestimated. Furthermore, BergBasic may not represent the continuing depletion of water storages.

When investigating the spatial distribution of the lag between TWS_{obs} and TWS_{mod}, a stronger disagreement gets obvious for the arid and boreal semi-arid regions, that hence are also characterized by lowest correlation values. Similar, Ngo-Duc et al. (2007), Swenson and Milly (2006) and Schmidt et al. (2008) received differences in the phase for arid regions as the deserts of Western China, Mongolia and South West North America. These regions are characterized by a low TWS amplitude, and thus the errors of the GRACE signal itself are high in relative terms (Güntner, 2008).

Except for the arid and semi-arid zone, the lag between TWS_{mod} and TWS_{obs} seems to be related to lower elevations and large river basins (Fig. 37). Accordingly, no lag is present in Scandinavia, the Central Siberian Plateau, East Siberia and Kamchatka, which tend to be areas of higher elevation. Intermediate lowlands as associated with the Ob and Lena basin, on contrary show a shift in the phase of TWS and thus also lower correlation in their time series. Here the preceding of TWS_{mod} may result from neglection of snow melt resp. runoff from adjacent uplands, that in reality with delay add water to the grid's TWS. Another reason may be a longer retardation time due to smaller slopes and the buffering effects of wetlands, lakes, and the river itself. Regarding high latitudes in North America, for similar reasons a lag is present for the Alaskan Yukon basin and the Canadian Mackenzie basin. A lag of two months in central Canada can be attributed to the numerous lakes and surface water bodies that characterize this region, yet are not represented in the model. For the somewhat higher elevations of the Labrador Peninsula, which lacks of large rivers, the MSC fits again. Indicative of the link to river retardation is the existence of numerous surface water bodies on the Labrador Peninsula, and the missing lag for the lowlands surrounding the Hudson Bay.

However, considering the whole study area, no statistical correlation between a lag in MSC and elevation, the occurrence of lakes and wetlands (data based on Lehner and Döll (2004)), or estimates of water holding capacity (data based on Webb et al. (2000)) is detectable. This suggests, that the mismatch cannot be ascribed to one cause alone, but different effects for different regions. While it is more likely that the lag in higher latitudes, resp. the polar and boreal zones, results from the effects of large rivers, the preceding in arid and boreal-semi arid zones may be caused by general poorer representation of low TWS variations, which are associated with a relative high error. Besides, inadequate parametrization using global uniform parameter values, as well as deficiencies in the meteorological forcing and errors in GRACE estimates influence the explained discrepancies.

Anomalies to the Mean Seasonal Cycle

In comparison to seasonal variations, the model clearly performs poorer in reproducing inter-annual variability. This usually is the case for hydrologic models, as a result of storage parametrization and missing processes (Landerer and Swenson, 2012; Zeng et al., 2008). Besides, also the observational data streams suffer from deficiencies and uncertainties in capturing long-term variations.

The fact that the average SWE_{obs} anomalies of the study area show less short term fluctuations than TWS_{obs} anomalies suggests that TWS_{obs} anomalies are dominated by other factors, as variations of summertime precipitation, groundwater fluctuations and change of surface water bodies, but also non-hydrologic effects such as seismic deformations (Güntner, 2008). In comparison, the modelled TWS anomalies are more smooth, and in their amplitude comparable to wintertime SWE_{mod}, while anomalies of W only pause through during the summer months. These variations in W are then dominated by variations in SM, while the amplitude of the indirectly estimated GW is extremely slight. This suggests that missing resp. insufficient representation of water retardation and extraction diminishes the model's ability to reproduce short term fluctuations in inter-annual variability. Regarding summertime variations, additional ET_{obs} is known to underestimate inter-annual variability (Jung et al., in preparation), and thus partly accounts for the insufficient representation of TWS fluctuations.

Nevertheless, the model reproduces the general tendencies of TWS anomalies. Thereby highest correlation is obtained for the continent centers and largest errors for the polar and arid zones. In the polar zone and high latitudes, one factor for these differences may be the neglection of the effects of melting permafrost, that in long-term changes TWS. Additionally, TWS_{obs} probably includes water changes caused by variations in (sea) ice and glaciers that due to the leakage error pass through (Güntner, 2008). So, a decreasing TWS_{obs} anomaly towards more negative values in the polar zone that cannot be simulated (see A 18) relates to the currently observed decrease in glacier and ice mass (IPCC, 2014). Further, Rangelova et al. (2007) showed that effects from post-glacial rebound may not be sufficiently removed, and influence e.g. TWS variations on the Labrador Peninsula. This is a region where the obtained correlation between TWS_{obs} and TWS_{mod} anomalies is low, too. In southern, and especially arid regions, an insufficient representation resp. parametrization of long-term storages as deep groundwater and the extraction of deep soil water for ET are among the main reasons for discrepancies. Further, the model does not consider human water withdrawal, which in dry years leads to a more pronounced negative anomaly of TWS (Güntner, 2008).

Besides issues due to model structure and leakage errors in TWS_{obs}, the gain factors, which are needed to reduce filtering effects in the GRACE TWS estimates, are optimized to recover variations on seasonal time scales. Thus uncertainty in their suitability for inter-annual and long-term interpretations exists (Landerer and Swenson, 2012).

6.4.2 Composition of TWS

One aim of this study is to determine the relative contributions of snow, soil moisture, and groundwater dynamics on variations of TWS, and to identify whether and how they vary temporally and spatially. Therefore, the intra-annual and inter-annual variability of TWS_{obs} and TWS_{mod} have been investigated and compared, and finally the dominant factor of variability in TWS_{mod} has been identified for both time-scales.

Mean Seasonal Cycle

In general, the seasonal variability of TWS_{obs} is reproduced well for most of the study area by the model. Exceptions are highest variations associated with an underestimation for the Canadian Archipelago and an overestimation for Scandinavia, for which already poorest performance regarding the MSC has been shown. As already discussed before, these divergences are likely due to leakage errors in the GRACE signal. Besides, underestimations of high TWS_{obs} variability may result from uncertainties in the meteorological forcing, as precipitation for example fails to record extreme events with high rainfall intensities.

The results show, that variations in the MSC of TWS in northern regions are mainly influenced by variations in SWE, and thus confirm earlier studies (Rangelova et al., 2007). On contrary, variations of W prevail south of approximately 50° latitudes, which in turn primarily results from variations in SM. This endorses the findings of Ngo-Duc et al. (2007), who showed that TWS in the Mississippi basin is mainly influenced by seasonal variations in soil moisture. The more northwards reaching influence of W in western Eurasia as well as the spatial pattern in North America goes along with the temperature gradient. Moreover, comparison with average annual temperature suggests, that for $T > 10$ °C variability of SM dominates, while for $T < 0$ °C snow dynamics prevail. Although this is plausible, it also highlights the dependency on the temperature data set used for model forcing.

The variability in the MSC of W is primarily caused by variations of SM. Simulated soil moisture is highly variable as it is influenced by various factors including changes in the amount of precipitation, snow melt, and evapotranspiration, as well as (constantly) released Qs. On contrary, GW, which is estimated indirectly from Qs and represents all delayed runoff, depends primarily on soil moisture and thus in general shows less variation. Nevertheless, the influence of GW prevails on the Canadian Arctic Archipelago, the northern coast of Eurasia, East Siberia and East China, which except for the Canadian Arctic Archipelago are snow dominated regions that receive relatively more annual precipitation. Here the soil is mostly saturated, due to snow melt in spring, sufficient precipitation in summer and low ET, and thus less variable. Additional high summer time precipitation then leads to increased Qs, of which the delayed amount in the BergBasic is ascribed to GW.

As the co-variance between SWE_{mod} and W is low, but not zero, uncertainty remains. Especially regions that experience deep snow conditions show a negative covariance, meaning high SWE_{mod} values are associated with low W. Considering the model structure, this is reasonable because when snow accumulates, no water infiltrates, which decreases SM as runoff still is possible. Vice versa, when snow melts or is not present, infiltration leads to high SM. This relation is more pronounced in regions, where much snow accumulates during winter and that experience high amounts of precipitation in summer.

In contrast to SWE_{mod} and W, the seasonal covariance between SM and GW is more pronounced and shows positive correlation for the coastal regions as well as for western Eurasia, that tend to receive more precipitation (ca. > 600 mm a^{-1}). In this context more precipitation leads to higher SM, ergo to higher Qs and thus to more delayed water that is ascribed to GW.

Anomalies to the Mean Seasonal Cycle

Regarding the inter-annual variability, the model performs poorer and does especially not reflect high TWS_{obs} fluctuations, as already shown. Reasons are again the insufficient representation of long-term water storages in the model, poor representation of inter-annual variability in ET_{obs} and other meteorological forcing, as well as errors in TWS_{obs} due to the processing of GRACE data (e.g. gain factors adjusted for seasonal variations) and leakage errors, that probably let the spatial distribution of the variability in TWS_{obs} anomalies appear patchier. As additional some co-variance between SWE_{mod} and W resp. SM and GW is present, predictions of the relative composition of inter-annual TWS variability are highly uncertain.

Nevertheless, the results show that in comparison to intra-annual variability, the inter-annual variations in TWS_{mod} are dominated by W. This is mainly caused by highly variable summertime precipitation events. On contrary, the results suggest that annual snow fall is less variable, as the variations in SWE_{mod} are less pronounced compared to W. Exceptions are regions of high maximum SWE, that accordingly show a considerable relative contribution of SWE_{mod} to the inter-annual TWS_{mod} variability.

Similar to the intra-annual variability, and for the same reasons, W is dominated by variations in SM, while GW solely shows higher contribution in regions of high SWE_{mod} amounts.

Résumé

In conclusion, the model is able to simulate seasonal variability of TWS_{mod} well, and reveals that in northern regions it is dominated by variations in SWE, while in lower latitudes soil moisture dynamics prevail. Inter-annual variations on contrary are mainly dominated by soil moisture, whereby the model showed poorer performance in representing variations in TWS_{obs}.

In general, model-based partitioning of TWS in different compartments is related to various uncertainties, as it depends on the model structure (e.g. which water storages are considered) and based on parametrization, the model gives divergent estimates (e.g. more soil moisture or

more groundwater) (Xie et al., 2012). However, as this study additional applied a snow product, and good resp. reasonable agreement has been obtained, confidence in the variability of the snow dynamics exists. Since further, intra-annual TWS variability is reproduced well, at least the relative contributions of SWE_{mod} and W to the seasonal variability of TWS_{mod} are expected to be reliable. Besides, some uncertainty remains regarding the inter-annual influence on TWS variations, as well as regarding the contribution of SM and GW to the variability in W, especially as GW in BergBasic is estimated indirectly and strongly affected by SM.

6.4.3 Trends in TWS

Besides general model performance, this study examines whether trends in GRACE TWS observations can be simulated by the model. Therefore, the Mann-Kendall test applying Sen's slope has been used. As this test is based on the median of the slopes between each value of the time series and its successor, it is not influenced by outliers, but prone to autocorrelation in the time series, which is a possible source of uncertainty in the estimated trends.

Similar to inter-annual variability and anomalies, the model in general tends to underestimate the observed trends. Qualitatively, the trend directions can be reproduced with an accuracy of 42 %, in equal shared by positive and negative directions. This is better than by chance, but still denotes a mismatch for almost 60 % of the study area. Thus, the findings show, that the estimation of inter-annual trends with the model is quite uncertain and corresponds to observations only in few regions of the study area. Additionally, the brevity of the considered time series allows few interpretation (Güntner, 2008). Moreover, as the gain factors are adjusted to seasonal interpretation, they "may thus not be suitable to quantify trends from GRACE Tellus land data" (Watkins et al., 2015).

Nevertheless, modelled and observed trends coincidence for some regions, for instance regarding negative trends in northern regions as East Siberia, Alaska and the Canadian Arctic Archipelago, that are associated with loss of snow and ice (Ditmar et al., 2010). Discrepancies in negative trends in these northern regions as in may result from leakage errors of adjacent melting sea ice or glaciers, or from not considered melting of permafrost. Fasullo et al. (2016) found negative trends in mid-latitudes as in the Eurasian Steppe to be caused by decreasing precipitation and drying due to natural variability as associated with the Atlantic Multi-Decadal Oscillation or the Pacific Decadal Oscillation and further teleconnections. These trends are also simulated by the model. Positive trends on contrary are in general less explicitly interpretable (Ditmar et al., 2010), but may also relate to large-scale oscillation feedbacks. While such positive trends in central North America and parts of Eurasia are simulated, the model fails to predict positive trends observed for the Labrador Peninsula. Here TWS_{obs} may partly suffer from insufficient correction of the post-glacial rebound, which causes a positive trend and is not considered in the model (Ditmar et al., 2010).

7 Conclusion and Outlook

The aim of this thesis was to develop a simple hydrological model that simulates the dynamics of the terrestrial water storage and its components in snow affected regions on the Northern Hemisphere, and to constrain it via multi-criteria calibration to GRACE TWS and GlobSnow SWE observations. Therefor various optimization methods have been tested to identify the most suitable for multi-criteria calibration. Further, different model formulations of soil moisture dynamics and runoff generation have been implemented and examined in order to determine the most appropriate model approach that achieves maximum consistency with observational data. Based on these findings, the calibrated 'best' model variant has been applied to investigate whether variations of the observed GRACE TWS can be reproduced and how the contributions of snow, soil moisture and groundwater dynamics to the TWS signal vary spatially and temporally.

Out of the tested optimization methods, the Covariance Matrix Adaption Evolution Strategy (CMAES) performed superior in terms of computational time and provided solution when compared to other search algorithm. Further, known data issues are considered best by CF6 that is composed of the wMEF of each constraint, which integrates uncertainty estimates of the observed variable, respectively. Thus, the combination of CMAES and CF6 is found to be the most suitable optimization method for multi-criteria calibration in this study.

The obtained differences of the implemented model variants in terms of spatial and temporal performance are minor, as the models' complexity is comparable and the calibrated parameter sets have similar effects. Also, incorporation of spatial distributed base flow parametrization yields no additional advantage to gain a better fit to observed TWS. However, the Bergström approach for soil dynamics achieved the best agreement with GRACE TWS data, whereby BergBasic is preferable to the groundwater variant as it showed slightly better performance and includes one parameter less.

In terms of spatial and temporal consistency to the observational data, BergBasic in general performs considerably well, whereby arid and semi-arid regions are identified as areas of main discrepancies, primarily related to insufficient parametrization and representation of relevant processes in these less snow-affected areas. Besides, problems in areas as the North American west coast and the Canadian Arctic Archipelago, where SWE_{obs} is moderate but TWS_{obs} is high, as well as a lag in the phase of the mean seasonal cycle between modelled and observed TWS suggest that inconsistencies between the GRACE and GlobSnow data streams exist, which in combination with limitations due to the simple model structure and errors in the input data lead to discrepancies.

Despite the preceding, BergBasic is able to capture the mean seasonal variability of observed TWS, and reveals that intra-annual variations of TWS in northern regions are dominated by variations in SWE, while in lower latitudes soil moisture dynamics prevail. Inter-annual variations on contrary are mainly dominated by soil moisture. Thereby the model showed poorer performance as it tends to underestimate variations of TWS_{obs}. Similar, trends in TWS anomalies are solely reproduced with an accuracy of 42 %, and are underestimated as well. The interpretation of both, inter-annual variations and trends, is further limited by not-negligible uncertainty in the input data and limitations in the model structure.

Albeit the findings of this study are promising, some issues remain and require further research. For one thing, calibration and comparison of the model variants revealed considerably problems due to over-parametrization, in particular of the snow component. Therefore, parameter correlations have to be inspected and sensitivity analysis is advisable to estimate the uncertainty introduced by the individual parameters. To further reduce parameter interdependencies, it is commendable to either neglect explicit simulation of sublimation from snow, to apply the traditional day-degree approach, or to better constrain the model e.g. by discharge observations. In this context, future efforts should not exclude a groundwater variant, as they have higher potential to simulate long-term retardation and the timing of runoff correctly, which probably becomes necessary when discharge is used as a constraint. Applying discharge as constraint may additionally help to better define retardation of runoff and thus improve the representation of the phase in TWS. In this context also inclusion of river routing can improve compliance, as the lag tends to relate to low elevations and large river basins. Moreover, the findings suggest that discrepancies in the TWS phase result from spatially varying causes and retardation effects with different times that cannot be captured by one global parameter. Thus, spatial varying of the delay factors ***q_t*** resp. ***g_d*** should be tested. In the present study, spatial distributed recession coefficients haven't improved model results in terms of costs, but the applied information probably was not consistent with the model structure. Besides the delay factor, spatial distributed values of ***p_et*** yield the potential to improve model performance, especially in arid and semi-arid regions.

In conclusion, this master thesis identified a simple hydrological model with 9 effective parameters that provides mean seasonal variations of TWS with substantially well conformity to GRACE TWS estimates for snow affected regions on the Northern Hemisphere. The model also allows determination of the influence of snow and soil moisture dynamic on these variations, yet its ability to predict inter-annual variability and trends is limited. Further efforts are necessary to better constrain the model and improve representation of spatial and temporal variability.

References

Abbott, M.B., Bathurst, J.C., Cunge, J.A., O'Connell, P.E., Rasmussen, J., 1986: An introduction to the European Hydrological System - Systeme Hydrologique Europeen,"SHE", 1: History and philosophy of a physically-based, distributed modelling system. Journal of Hydrology, 87(1): 45-59.

AG, Bodenkunde, 2005: Bodenkundliche Kartieranleitung (KA5), 5 Schweitzerbart'sche Verlagsbuchhandlung, Stuttgart.

Alcamo, J., Döll, P., Henrichs, T., Kaspar, F., Lehner, B., Rösch, T., Siebert, S., 2003: Development and testing of the WaterGAP 2 global model of water use and availability. Hydrological Sciences Journal, 48(3): 317-337.

Alkama, R., Decharme, B., Douville, H., Becker, M., Cazenave, A., Sheffield, J., Voldoire, A., Tyteca, S., Le Moigne, P., 2010: Global evaluation of the ISBA-TRIP continental hydrological system. Part I: Comparison to GRACE terrestrial water storage estimates and in situ river discharges. Journal of Hydrometeorology, 11(3): 583-600.

Alley, W.M., 1984: On the treatment of evapotranspiration, soil moisture accounting, and aquifer recharge in monthly water balance models. Water Resources Research, 20(8): 1137-1149.

Arnell, N.W., 1999: A simple water balance model for the simulation of streamflow over a large geographic domain. Journal of Hydrology, 217(3): 314-335.

Arsenault, R., Poulin, A., Côté, P., Brissette, F., 2013: Comparison of stochastic optimization algorithms in hydrological model calibration. Journal of Hydrologic Engineering, 19(7): 1374-1384.

Balsamo, G., Beljaars, A., Scipal, K., Viterbo, P., van den Hurk, B., Hirschi, M., Betts, A.K., 2009: A revised hydrology for the ECMWF model: Verification from field site to terrestrial water storage and impact in the Integrated Forecast System. Journal of Hydrometeorology, 10(3): 623-643.

Battin, T.J., Luyssaert, S., Kaplan, L.A., Aufdenkampe, A.K., Richter, A., Tranvik, L.J., 2009: The boundless carbon cycle. Nature Geoscience, 2(9): 598-600.

Beck, C., Grieser, J., Rudolf, B., Schneider, U., 2005: A new monthly precipitation climatology for the global land areas for the period 1951 to 2000, Geophysical Research Abstracts, pp. 07154.

Beck, H.E., De Roo, A., van Dijk, A.I., 2015: Global Maps of Streamflow Characteristics Based on Observations from Several Thousand Catchments. Journal of Hydrometeorology, 16(4): 1478-1501.

Beck, H.E., Dijk, A.I., Miralles, D.G., Jeu, R.A., McVicar, T.R., Schellekens, J., 2013: Global patterns in base flow index and recession based on streamflow observations from 3394 catchments. Water Resources Research, 49(12): 7843-7863.

Behrangi, A., Khakbaz, B., Jaw, T.C., AghaKouchak, A., Hsu, K., Sorooshian, S., 2011: Hydrologic evaluation of satellite precipitation products over a mid-size basin. Journal of Hydrology, 397(3): 225-237.

Bergström, S., 1976: Development and application of a conceptual runoff model for Scandinavian catchments, Swedish Meteorological and Hydrological Institite (SMHI), Norrköping, Sweden.

Bergström, S., 1995: The HBV model. In: Singh, V. (Ed.), Computer models of watershed hydrology., pp. 443-476.

Beven, K., 2006: A manifesto for the equifinality thesis. Journal of Hydrology, 320(1): 18-36.

Beven, K., Kirkby, M.J., 1979: A physically based, variable contributing area model of basin hydrology/Un modèle à base physique de zone d'appel variable de l'hydrologie du bassin versant. Hydrological Sciences Journal, 24(1): 43-69.

Beven, K., Lamb, R., Quinn, P., Romanowicz, R., Freer, J., 1995: Topmodel. In: Singh, V. (Ed.), Computer models of watershed hydrology, pp. 627-668.

Blume, H.-P., Brümmer, G.W., Horn, R., Kandeler, E., Kögel-Knabner, I., Kretzschmar, R., Stahr, K., Wilke, B.-M., Thiele-Bruhn, S., Welp, G., 2010: Scheffer/Schachtschabel: Lehrbuch der Bodenkunde, 16. Spektrum Akademischer Verlag, Heidelberg, Germany.

Bondeau, A., Smith, P.C., Zaehle, S., Schaphoff, S., Lucht, W., Cramer, W., Gerten, D., Lotze-Campen, H., Müller, C., Reichstein, M., 2007: Modelling the role of agriculture for the 20th century global terrestrial carbon balance. Global Change Biology, 13(3): 679-706.

Brooks, K.N., Ffolliott, P.F., Magner, J.A., 2012: Hydrology and the Management of Watersheds, 4. Wiley-Blackwell, Oxford, UK.

Broyden, C.G., 1970: The convergence of a class of double-rank minimization algorithms 1. general considerations. IMA Journal of Applied Mathematics, 6(1): 76-90.

Brutsaert, W., Nieber, J.L., 1977: Regionalized drought flow hydrographs from a mature glaciated plateau. Water Resour. Res, 13(3): 637-643.

Budyko, M.I., 1961: The heat balance of the earth's surface. Soviet Geography, 2(4): 3-13.

Burnash, R., Singh, V., 1995: The NWS river forecast system-Catchment modeling. Computer models of watershed hydrology.: 311-366.

Byrd, R.H., Gilbert, J.C., Nocedal, J., 2000: A trust region method based on interior point techniques for nonlinear programming. Mathematical Programming, 89(1): 149-185.

Climate Service Center Germany, G., 2012: Klimawissen Verdunstung. <http://www.climate-service-center.de/033633/index_0033633.html.de#spez>. (Last Update: 31.07.2012), (Access: 20.03.2016).

Coello, C.C., Lamont, G.B., Van Veldhuizen, D.A., 2007: Evolutionary algorithms for solving multi-objective problems. Springer.

Crow, W.T., 2007: A novel method for quantifying value in spaceborne soil moisture retrievals. Journal of Hydrometeorology, 8(1): 56-67.

Deardorff, J., 1978: Efficient prediction of ground surface temperature and moisture, with inclusion of a layer of vegetation. Journal of Geophysical Research: Oceans, 83(C4): 1889-1903.

Deb, K., Pratap, A., Agarwal, S., Meyarivan, T., 2002: A fast and elitist multiobjective genetic algorithm: NSGA-II. IEEE transactions on Evolutionary Computation, 6(2): 182-197.

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

Derksen, C., Lemmetyinen, J., Toose, P., Silis, A., Pulliainen, J., Sturm, M., 2014: Physical properties of Arctic versus subarctic snow: Implications for high latitude passive microwave snow water equivalent retrievals. Journal of Geophysical Research: Atmospheres, 119(12): 7254-7270.

Devia, G.K., Ganasri, B., Dwarakish, G., 2015: A Review on Hydrological Models. Aquatic Procedia, 4: 1001-1007.

Dickinson, R.E., 1986: Biosphere/atmosphere transfer scheme (BATS) for the NCAR community climate model. Technical report.

Ditmar, P., Liu, X., Klees, R., Revtova, E., Vermeersen, B., Riva, R., Siemes, C., Zhao, Q., 2010: Regular Gravity Field Variations and Mass Transport in the Earth System from DEOS Models Based on GRACE Satellite Data. Gravity, Geoid and Earth Observation: IAG Commission 2: Gravity Field, Chania, Crete, Greece, 23-27 June 2008, 135. Springer, Berlin.

Döll, P., Douville, H., Güntner, A., Schmied, H.M., Wada, Y., 2015: Modelling freshwater resources at the global scale: challenges and prospects. Surveys in Geophysics: 1-27.

Döll, P., Fritsche, M., Eicker, A., Schmied, H.M., 2014: Seasonal water storage variations as impacted by water abstractions: comparing the output of a global hydrological model with GRACE and GPS observations. Surveys in Geophysics, 35(6): 1311-1331.

Döll, P., Kaspar, F., Lehner, B., 2003: A global hydrological model for deriving water availability indicators: model tuning and validation. Journal of Hydrology, 270(1): 105-134.

Dümenil, L., Todini, E., 1992: A rainfall-runoff scheme for use in the Hamburg climate model. Advances in Theoretical Hydrology, A Tribune to James Dooge: 129-157.

ECMWF, 2014: IFS Documentation - Cy40r1: Part IV: Physical Processes. European Centre for Medium-Range Weather Forecasts, Reading, England.

Evans, J., Perlman, H., 2015: The Water Caycle. U.S. Department of the Interioir, U.S. Geological Survey.

Falkowski, P., Scholes, R., Boyle, E., Canadell, J., Canfield, D., Elser, J., Gruber, N., Hibbard, K., Högberg, P., Linder, S., 2000: The global carbon cycle: a test of our knowledge of earth as a system. Science, 290(5490): 291-296.

Fasullo, J., Lawrence, D., Swenson, S., 2016: Are GRACE-era Terrestrial Water Trends Driven by Anthropogenic Climate Change? Advances in Meteorology, 2016.

Feng, W., Zhong, M., Lemoine, J.M., Biancale, R., Hsu, H.T., Xia, J., 2013: Evaluation of groundwater depletion in North China using the Gravity Recovery and Climate Experiment (GRACE) data and ground-based measurements. Water Resources Research, 49(4): 2110-2118.

Fischer, C., 2013: Automatische Kalibrierung hydrologischer Modelle. Entwicklung und Anwendung des Kalibrierungssystems OPTAS, Friedrich-Schiller University, Jena.

Flügel, W.A., 1995: Delineating hydrological response units by geographical information system analyses for regional hydrological modelling using PRMS/MMS in the drainage basin of the River Bröl, Germany. Hydrological Processes, 9(3-4): 423-436.

Foglia, L., Hill, M., Mehl, S., Burlando, P., 2009: Sensitivity analysis, calibration, and testing of a distributed hydrological model using error-based weighting and one objective function. Water Resources Research, 45(6).

Forman, B.A., Reichle, R., Rodell, M., 2012: Assimilation of terrestrial water storage from GRACE in a snow-dominated basin. Water Resources Research, 48(1).

Friedl, M.A., Sulla-Menashe, D., Tan, B., Schneider, A., Ramankutty, N., Sibley, A., Huang, X., 2010: MODIS Collection 5 global land cover: Algorithm refinements and characterization of new datasets. Remote Sensing of Environment, 114(1): 168-182.

Fu, B.P., 1981: On the calculation of the evaporation from land surface [in Chinese]. Sci. Atmos. Sin, 5(1): 23-31.

Gao, H., Tang, Q., Shi, X., Zhu, C., Bohn, T., Su, F., Sheffield, J., Pan, M., Lettenmaier, D., Wood, E.F., 2010: Water budget record from Variable Infiltration Capacity (VIC) model. Algorithm Theoretical Basis Document for Terrestrial Water Cycle Data Records.

Gerten, D., Schaphoff, S., Haberlandt, U., Lucht, W., Sitch, S., 2004: Terrestrial vegetation and water balance—hydrological evaluation of a dynamic global vegetation model. Journal of Hydrology, 286(1): 249-270.

Gulden, L.E., Rosero, E., Yang, Z.L., Wagener, T., Niu, G.Y., 2008: Model performance, model robustness, and model fitness scores: A new method for identifying good land-surface models. Geophysical Research Letters, 35(11).

Güntner, A., 2008: Improvement of global hydrological models using GRACE data. Surveys in Geophysics, 29(4-5): 375-397.

Guo, Z., Dirmeyer, P.A., Hu, Z.Z., Gao, X., Zhao, M., 2006: Evaluation of the Second Global Soil Wetness Project soil moisture simulations: 2. Sensitivity to external meteorological forcing. Journal of Geophysical Research: Atmospheres, 111(D22).

Gupta, H.V., Beven, K.J., Wagener, T., 2005: Model calibration and uncertainty estimation. Encyclopedia of Hydrological Sciences.

Gupta, H.V., Kling, H., Yilmaz, K.K., Martinez, G.F., 2009: Decomposition of the mean squared error and NSE performance criteria: Implications for improving hydrological modelling. Journal of Hydrology, 377(1): 80-91.

Gupta, S.K., 2011: Modern hydrology and sustainable water development. Wiley-Blackwell, Oxford, UK.

Gustard, A., Bullock, A., Dixon, J., 1992: Low flow estimation in the United Kingdom. Institute of Hydrology.

Haddeland, I., Clark, D.B., Franssen, W., Ludwig, F., Voß, F., Arnell, N.W., Bertrand, N., Best, M., Folwell, S., Gerten, D., 2011: Multimodel estimate of the global terrestrial water balance: Setup and first results. Journal of Hydrometeorology, 12(5): 869-884.

Hagemann, S., 2011: The hydrological cycle: how observational data are able to improve climate models, Hamburg University Hamburg.

Hagemann, S., Dümenil, L., 1997: A parametrization of the lateral waterflow for the global scale. Climate Dynamics, 14(1): 17-31.

Hagemann, S., Gates, L.D., 2003: Improving a subgrid runoff parameterization scheme for climate models by the use of high resolution data derived from satellite observations. Climate Dynamics, 21(3-4): 349-359.

Hamon, W.R., 1963: Computation of direct runoff amounts from storm rainfall. Int. Assoc. Sci. Hydrol. Pub., 63: 52-63.

Hanasaki, N., Kanae, S., Oki, T., Masuda, K., Motoya, K., Shirakawa, N., Shen, Y., Tanaka, K., 2008: An integrated model for the assessment of global water resourcesâ€“Part 1: Model description and input meteorological forcing. Hydrology and Earth System Sciences, 12(4): 1007-1025.

Hancock, S., Baxter, R., Evans, J., Huntley, B., 2013: Evaluating global snow water equivalent products for testing land surface models. Remote Sensing of Environment, 128: 107-117.

Hansen, N., 2014: The CMA Evolution Strategy. <https://www.lri.fr/~hansen/cmaesintro.html>. (Last Update: 16.03.2014), (Access: 30.03.2016).

Hansen, N., Kern, S., 2004: Evaluating the CMA evolution strategy on multimodal test functions, Parallel problem solving from nature-PPSN VIII. Springer, pp. 282-291.

Harding, R., Best, M., Blyth, E., Hagemann, S., Kabat, P., Tallaksen, L.M., Warnaars, T., Wiberg, D., Weedon, G.P., Lanen, H.v., 2011: WATCH: Current knowledge of the terrestrial global water cycle. Journal of Hydrometeorology, 12(6): 1149-1156.

Hari, P., Petäjä, T., Bäck, J., Kerminen, V.-M., Lappalainen, H.K., Vihma, T., Laurila, T., Viisanen, Y., Vesala, T., Kulmala, M., 2016: Conceptual design of a measurement network of the global change. Atmospheric Chemistry and Physics, 16(2): 1017-1028.

He, Z., Parajka, J., Tian, F., Blöschl, G., 2014: Estimating degree-day factors from MODIS for snowmelt runoff modeling. Hydrology and Earth System Sciences, 18(12): 4773-4789.

Huffman, G.J., Adler, R.F., Morrissey, M.M., Bolvin, D.T., Curtis, S., Joyce, R., McGavock, B., Susskind, J., 2001: Global precipitation at one-degree daily resolution from multisatellite observations. Journal of Hydrometeorology, 2(1): 36-50.

Huffman, G.J., Bolvin, D.T., 2013: GPCP Version 2.2 SG Combined Precipitation Data Set Documentation. NASA Goodard Space Flight Center, Greenbelt, MD, USA.

Immerzeel, W., Droogers, P., 2008: Calibration of a distributed hydrological model based on satellite evapotranspiration. Journal of Hydrology, 349(3): 411-424.

IPCC, 2014: Climate change 2014: Synthesis Report. Contribution of working groups I, II and III to the fifth assessment report of the intergovernmental panel on climate change. IPCC, Geneva, Switzerland, 151 pp.

Ivanov, V.Y., Vivoni, E.R., Bras, R.L., Entekhabi, D., 2004: Catchment hydrologic response with a fully distributed triangulated irregular network model. Water Resources Research, 40(11).

Jakeman, A., Hornberger, G., 1993: How much complexity is warranted in a rainfall-runoff model? Water Resources Research, 29(8): 2637-2649.

Jones, P.D., Harris, I., 2008: Climatic Research Unit (CRU) time-series datasets of variations in climate with variations in other phenomena. <http://catalogue.ceda.ac.uk/uuid/3f8944800cc48e1cbc29a5ee12d8542d>. (Last Update: 2015), (Access: 06.01.2016).

Julander, R., 2000: Degree Day Models. <http://www.civil.utah.edu/~cv5450/modeling/>. (Last Update: 20.01.2000), (Access: 23.04.20162016).

Jung, M., Zscheischler, J., 2013: A guided hybrid genetic algorithm for feature selection with expensive cost functions. Procedia Computer Science, 18: 2337-2346.

Kalnay, E., Kanamitsu, M., Kistler, R., Collins, W., Deaven, D., Gandin, L., Iredell, M., Saha, S., White, G., Woollen, J., 1996: The NCEP/NCAR 40-year reanalysis project. Bulletin of the American meteorological Society, 77(3): 437-471.

Kendall, M., 1975: Rank Correlation Methods. Griffin & Co, London.

Kim, H., Yeh, P.J.F., Oki, T., Kanae, S., 2009: Role of rivers in the seasonal variations of terrestrial water storage over global basins. Geophysical Research Letters, 36(17).

Kirchner, J.W., 2009: Catchments as simple dynamical systems: Catchment characterization, rainfall-runoff modeling, and doing hydrology backward. Water Resources Research, 45(2).

Kirkpatrick, S., Gelatt, C.D., Vecchi, M.P., 1983: Optimization by simmulated annealing. Science, 220(4598): 671-680.

Klemeš, V., 1986: Operational testing of hydrological simulation models. Hydrological Sciences Journal, 31(1): 13-24.

Köppen, W., 1900: Versuch einer Klassifikation der Klimate, vorzugsweise nach ihren Beziehungen zur Pflanzenwelt. Geographische Zeitschrift, 6(11. H): 593-611.

Korzoun, V.I., Sokolev, A.A., Budyko, M.I., Voskresensky, K.P., LKalinin, G.P., Konoplyansev, A.A., Korotkevich, E.S., Lvovich, M.I., 1978: Atlas of World Water Balance. UNESCO, Paris, France.

Koster, R.D., P. Mahanama, S.P., 2012: Land surface controls on hydroclimatic means and variability. Journal of Hydrometeorology, 13(5): 1604-1620.

Kottek, M., Grieser, J., Beck, C., Rudolf, B., Rubel, F., 2006: World map of the Köppen-Geiger climate classification updated. Meteorologische Zeitschrift, 15(3): 259-263.

Krause, J., 2007: Parameteroptimierung bei der evolutionären Entwicklung neuronaler Topologien, Christian-Albrechts-Universität, Kiel.

Krause, P., Boyle, D., Bäse, F., 2005: Comparison of different efficiency criteria for hydrological model assessment. Advances in Geosciences, 5: 89-97.

Kruizinga, A., Williams, S., 2010: GRACE - Gravity Recovery and Climate Experiment. <http://www.csr.utexas.edu/grace/overview.html>. (Last Update: 20.08.2010), (Access: 13.01.2016).

Kustas, W.P., Rango, A., Uijlenhoet, R., 1994: A simple energy budget algorithm for the snowmelt runoff model. Water Resources Research, 30(5): 1515-1527.

Landerer, F., Swenson, S., 2012: Accuracy of scaled GRACE terrestrial water storage estimates. Water Resources Research, 48(4).

Legates, D.R., McCabe, G.J., 1999: Evaluating the use of “goodness-of-fit” measures in hydrologic and hydroclimatic model validation. Water resources research, 35(1): 233-241.

Lehn, J., Wegmann, H., 2006: Einführung in die Statistik. Springer, Berlin.

Lehner, B., Döll, P., 2004: Global Lakes and Wetlands database, Lakes and Wetlands Grid (Level 3) Dataset. <http://www.worldwildlife.org/publications/global-lakes-and-wetlands-database-lakes-and-wetlands-grid-level-3>.

Lettenmaier, D.P., Alsdorf, D., Dozier, J., Huffman, G.J., Pan, M., Wood, E.F., 2015: Inroads of remote sensing into hydrologic science during the WRR era. Water Resources Research, 51(9): 7309-7342.

Lindström, G., Johansson, B., Persson, M., Gardelin, M., Bergström, S., 1997: Development and test of the distributed HBV-96 hydrological model. Journal of Hydrology, 201(1): 272-288.

Liu, J., Li, Z., Huang, L., Tian, B., 2014: Hemispheric-scale comparison of monthly passive microwave snow water equivalent products. Journal of Applied Remote Sensing, 8(1): 084688-084688.

Livneh, B., Lettenmaier, D., 2012: Multi-criteria parameter estimation for the unified land model. Hydrology and Earth System Sciences, 16(8): 3029-3048.

Luojous, K., Pulliainen, J., Takala, M., Lemmetyinen, J., Kangwa, M., Eskelinen, M., Metsämäki, S., Solberg, R., Salberg, A.-B., Bippus, G., Ripper, E., Nagler, T., Derksen, C., Wiesmann, A., Wunderle, S., Hüsler, F., Fontana, F., Foppa, N., 2014: GlobSnow-2 Final Report, European Space Agency.

Manabe, S., 1969: Climate and the Ocean Circulation 1: I. The Atmospheric Circulation and the Hydrology of the Earth's Surface. Monthly Weather Review, 97(11): 739-774.

Mann, H.B., 1945: Nonparametric tests against trend. Econometrica: Journal of the Econometric Society: 245-259.

Marchenko, S., Romanovsky, V., Tipenko, G., 2008: Numerical modeling of spatial permafrost dynamics in Alaska, Proceedings of the ninth international conference on

permafrost. Institute of Northern Engineering, University of Alaska Fairbanks, pp. 1125-1130.

Marquardt, D.W., 1963: An algorithm for least-squares estimation of nonlinear parameters. Journal of the society for Industrial and Applied Mathematics, 11(2): 431-441.

MathWorks, I., 2015a: Global Optimization Toolbox User's Guide (R2015a). Mathworks Natick, MA, USA.

MathWorks, I., 2015b: Optimization Toolbox User's Guide (R2015a). Mathworks Natick, MA, USA.

Meigh, J., McKenzie, A., Sene, K., 1999: A grid-based approach to water scarcity estimates for eastern and southern Africa. Water Resources Management, 13(2): 85-115.

Melsen, L., Teuling, A., Berkum, S., Torfs, P., Uijlenhoet, R., 2014: Catchments as simple dynamical systems: A case study on methods and data requirements for parameter identification. Water Resources Research, 50(7): 5577-5596.

Milly, P., Shmakin, A., 2002: Global modeling of land water and energy balances. Part I: The land dynamics (LaD) model. Journal of Hydrometeorology, 3(3): 283-299.

Milzow, C., Krogh, P.E., Bauer-Gottwein, P., 2011: Combining satellite radar altimetry, SAR surface soil moisture and GRACE total storage changes for hydrological model calibration in a large poorly gauged catchment. Hydrology and Earth System Sciences, 15(6): 1729-1743.

Miralles, D., Holmes, T., De Jeu, R., Gash, J., Meesters, A., Dolman, A., 2011: Global land-surface evaporation estimated from satellite-based observations. Hydrology and Earth System Sciences, 15(2): 453-469.

Mitchell, T.D., Jones, P.D., 2005: An improved method of constructing a database of monthly climate observations and associated high-resolution grids. International Journal of Climatology, 25(6): 693-712.

Monteith, J., 1965: Evaporation and environment, Symp. Soc. Exp. Biol, pp. 4.

Moriasi, D., Arnold, J., Van Liew, M., Bingner, R., Harmel, R., Veith, T., 2007: Model evaluation guidelines for systematic quantification of accuracy in watershed simulations. Trans. Asabe, 50(3): 885-900.

MPI-BGC, 2016: Institute Profile. <https://www.bgc-jena.mpg.de/index.php/Institute/Profile>. (Last Update: 2016), (Access: 01.07.2016).

Müller-Schmied, H., Eisner, S., Franz, D., Wattenbach, M., Portmann, F.T., Florke, M., Doll, P., 2014: Sensitivity of simulated global-scale freshwater fluxes and storages to input data, hydrological model structure, human water use and calibration. Hydrology and Earth System Sciences, 18(9): 3511-3538.

Murphy, A.H., 1988: Skill scores based on the mean square error and their relationships to the correlation coefficient. Monthly Weather Review, 116(12): 2417-2424.

Murphy, D., Koop, T., 2005: Review of the vapour pressures of ice and supercooled water for atmospheric applications. Quarterly Journal of the Royal Meteorological Society, 131(608): 1539-1565.

NASA, C.T., 2013: CERES_SYN1deg_Ed3A Data Quality Summary. <http://ceres.larc.nasa.gov/documents/DQ_summaries/CERES_SYN1deg_Ed3A_DQS.pdf>. (Last Update: 08.08.2013), (Access: 08.01.2015).

Nash, J.E., Sutcliffe, J.V., 1970: River flow forecasting through conceptual models part I—A discussion of principles. Journal of Hydrology, 10(3): 282-290.

Neitsch, S., Arnold, J., Kiniry, J., Williams, J., King, K., 2002: SWAT Manual. USDA. Agricultural Research Service and Blackland Research Centre, Texas A&M University, USA.

New, M., Hulme, M., Jones, P., 2000: Representing twentieth-century space-time climate variability. Part II: Development of 1901-96 monthly grids of terrestrial surface climate. Journal of Climate, 13(13): 2217-2238.

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

Nijssen, B., Schnur, R., Lettenmaier, D.P., 2001: Global retrospective estimation of soil moisture using the variable infiltration capacity land surface model. Journal of Climate, 14(8): 1790-1808.

Niu, G.Y., Seo, K.W., Yang, Z.L., Wilson, C., Su, H., Chen, J., Rodell, M., 2007: Retrieving snow mass from GRACE terrestrial water storage change with a land surface model. Geophysical Research Letters, 34(15).

Orth, R., Koster, R.D., Seneviratne, S.I., 2013: Inferring soil moisture memory from streamflow observations using a simple water balance model. Journal of Hydrometeorology, 14(6): 1773-1790.

Pagano, T.C., Sorooshian, S., 2006: Global water cycle (fundamental, theory, mechanisms). Encyclopedia of Hydrological Sciences, 5. Wiley, Chichester, England.

Pan, X., Liu, Y., Fan, X., 2015: Comparative Assessment of Satellite-Retrieved Surface Net Radiation: An Examination on CERES and SRB Datasets in China. Remote Sensing, 7(4): 4899-4918.

Parajka, J., Blöschl, G., 2008: The value of MODIS snow cover data in validating and calibrating conceptual hydrologic models. Journal of Hydrology, 358(3): 240-258.

Penman, H.L., 1948: Natural evaporation from open water, bare soil and grass, Proceedings of the Royal Society of London A: Mathematical, Physical and Engineering Sciences. The Royal Society, pp. 120-145.

Pettyjohn, W.A., Henning, R., 1979: Preliminary estimate of ground-water recharge rates, related streamflow and water quality in Ohio. Ohio State University Water Resources Center Project Completion Report, 552.

Priestley, C., Taylor, R., 1972: On the assessment of surface heat flux and evaporation using large-scale parameters. Monthly Weather Review, 100(2): 81-92.

Pulliainen, J.T., Grandell, J., Hallikainen, M.T., 1999: HUT snow emission model and its applicability to snow water equivalent retrieval. Geoscience and Remote Sensing, IEEE Transactions on, 37(3): 1378-1390.

Ramillien, G., Frappart, F., Güntner, A., Ngo-Duc, T., Cazenave, A., Laval, K., 2006: Time variations of the regional evapotranspiration rate from Gravity Recovery and Climate Experiment (GRACE) satellite gravimetry. Water Resources Research, 42(10).

Rangelova, E., Van der Wal, W., Braun, A., Sideris, M., Wu, P., 2007: Analysis of Gravity Recovery and Climate Experiment time-variable mass redistribution signals over North America by means of principal component analysis. Journal of Geophysical Research: Earth Surface, 112(F3).

Rango, A., Martinec, J., 1995: Revisting the degree-day methd for snow melt computations. Journal of the American Water Resources Association, 31(4): 657-669.

Rees, H., Holmes, M., Young, A., Kansakar, S., 2004: Recession-based hydrological models for estimating low flows in ungauged catchments in the Himalayas. Hydrology and Earth System Sciences Discussions, 8(5): 891-902.

Refsgaard, J.C., 1996: Terminology, Modelling Protocol and Classification of Hydrological Model Codes. In: Abbott, M.B., Refsgaard, J.C. (Eds.), Distributed Hydrological Modelling. Kluwer Academic Publishers, Dordrecht, pp. 17-41.

Riebeek, H., 2011: The Carbon Cycle. <http://earthobservatory.nasa.gov/Features/CarbonCycle/page1.php>. (Last Update: 16.06.2011), (Access: 01.07.2016).

Rodell, M., Famiglietti, J., Chen, J., Seneviratne, S., Viterbo, P., Holl, S., Wilson, C., 2004: Basin scale estimates of evapotranspiration using GRACE and other observations. Geophysical Research Letters, 31(20).

Rodell, M., Velicogna, I., Famiglietti, J.S., 2009: Satellite-based estimates of groundwater depletion in India. Nature, 460(7258): 999-1002.

Rudolf, B., Rubel, F., 2005: Global Precipitation. In: Hantel, M. (Ed.), Observed Global Climate Geophysics Springer, Berlin, pp. 567.

Rust, H.W., Kruschke, T., Dobler, A., Fischer, M., Ulbrich, U., 2015: Discontinuous Daily Temperatures in the WATCH Forcing Datasets. Journal of Hydrometeorology, 16(1): 465-472.

Sai Krishna, S., Manavalan, P., Rao, P., 2014: Estimation of Net Radiation using satellite based data inputs. ISPRS-International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences, 1: 307-313.

Sakumura, C., Bettadpur, S., Bruinsma, S., 2014: Ensemble prediction and intercomparison analysis of GRACE time-variable gravity field models. Geophysical Research Letters, 41(5): 1389-1397.

Scheiter, S., Langan, L., Higgins, S.I., 2013: Next-generation dynamic global vegetation models: learning from community ecology. New Phytologist, 198(3): 957-969.

Schmidt, R., Flechtner, F., Meyer, U., Neumayer, K.-H., Dahle, C., König, R., Kusche, J., 2008: Hydrological signals observed by the GRACE satellites. Surveys in Geophysics, 29(4-5): 319-334.

Schneider, U., Fuchs, T., Meyer-Christoffer, A., Rudolf, B., 2015: Global precipitation analysis products of the GPCC>. (Last Update: 2015), (Access: 03.02.2016).

Sen, P.K., 1968: Estimates of the regression coefficient based on Kendall's tau. Journal of the American Statistical Association, 63(324): 1379-1389.

Seo, K.W., Ryu, D., Kim, B.M., Waliser, D.E., Tian, B., Eom, J., 2010: GRACE and AMSR-E-based estimates of winter season solid precipitation accumulation in the Arctic drainage region. Journal of Geophysical Research: Atmospheres, 115(D20).

Sloto, R.A., Crouse, M.Y., 1996: HYSEP, a computer program for streamflow hydrograph separation and analysis. US Department of the Interior, US Geological Survey.

Soliman, S.A.-H., Mantawy, A.-A.H., 2011: Modern optimization techniques with applications in electric power systems. Springer, Berlin.

Sood, A., Smakhtin, V., 2015: Global hydrological models: a review. Hydrological Sciences Journal, 60(4): 549-565.

Sorooshian, S., Duan, Q., Gupta, V.K., 1993: Calibration of rainfall-runoff models: application of global optimization to the Sacramento soil moisture accounting model. Water Resources Research, 29(4): 1185-1194.

Stacke, T., Claussen, M., Hagemann, S., 2011: Development of a dynamical wetlands hydrology scheme and its application under different climate conditions, University of Hamburg Hamburg.

Swenson, S.C., Milly, P., 2006: Climate model biases in seasonality of continental water storage revealed by satellite gravimetry. Water Resources Research, 42(3).

Syed, T.H., Famiglietti, J.S., Chambers, D.P., 2009: GRACE-based estimates of terrestrial freshwater discharge from basin to continental scales. Journal of Hydrometeorology, 10(1): 22-40.

Syed, T.H., Famiglietti, J.S., Chen, J., Rodell, M., Seneviratne, S.I., Viterbo, P., Wilson, C.R., 2005: Total basin discharge for the Amazon and Mississippi River basins from GRACE and a land-atmosphere water balance. Geophysical Research Letters, 32(24).

Takala, M., Luojus, K., Pulliainen, J., Derksen, C., Lemmetyinen, J., Kärnä, J.-P., Koskinen, J., Bojkov, B., 2011: Estimating northern hemisphere snow water equivalent for climate research through assimilation of space-borne radiometer data and ground-based measurements. Remote Sensing of Environment, 115(12): 3517-3529.

Tallaksen, L.M., Burkhart, J.F., Stordal, F., 2015: Land-Atmosphere Interactions in Cold Environments (LATICE): The role of Atmosphere-Biosphere-Cryosphere-Hydrosphere interactions in a changing climate, EGU General Assembly Conference Abstracts, pp. 11350.

Teuling, A., Lehner, I., Kirchner, J., Seneviratne, S., 2010: Catchments as simple dynamical systems: Experience from a Swiss prealpine catchment. Water Resources Research, 46(10).

Thomas Jr, H.A., 1981: Improved methods for national water assessment, water resources contract: WR15249270, Harvard Water Resources Group.

Thornthwaite, C.W., 1948: An approach toward a rational classification of climate. Geographical review, 38(1): 55-94.

Todini, E., 1996: The ARNO rainfall-runoff model. Journal of Hydrology, 175(1): 339-382.

Tramontana, G., Jung, M., Camps-Valls, G., Ichii, K., Raduly, B., Reichstein, M., Schwalm, C.R., Arain, M.A., Cescatti, A., Kiely, G., Merbold, L., Serrano-Ortiz, P., Sickert, S., Wolf, S., Papale, D., 2016: Predicting carbon dioxide and energy fluxes across global FLUXNET sites with regression algrithms. Biogeosciences Discussions, in review.

Trishchenko, A., 2002: Removing unwanted fluctuations in the AVHRR thermal calibration data using robust techniques. Journal of Atmospheric and Oceanic Technology, 19(12): 1939-1954.

van Beek, L., Bierkens, M.F., 2008: The Global Hydrological Model PCR-GLOBWB: Conceptualization, Parametrizaton and Verification, Department of Physical Geography, Utrecht University, Utrecht, The Netherlands.

Van Dijk, A., 2010: Climate and terrain factors explaining streamflow response and recession in Australian catchments. Hydrology and Earth System Sciences, 14(1): 159-169.

Vivoy, N., 2015: CRU-NCEP v6.1 Dataset. <http://dods.extra.cea.fr/data/p529viov/cruncep/>.

Vörösmarty, C.J., Federer, C.A., Schloss, A.L., 1998: Potential evaporation functions compared on US watersheds: Possible implications for global-scale water balance and terrestrial ecosystem modeling. Journal of Hydrology, 207(3): 147-169.

Vörösmarty, C.J., Moore, B., Grace, A.L., Gildea, M.P., Melillo, J.M., Peterson, B.J., Rastetter, E.B., Steudler, P.A., 1989: Continental scale models of water balance and fluvial transport: an application to South America. Global biogeochemical cycles, 3(3): 241-265.

Vrugt, J.A., Gupta, H.V., Bastidas, L.A., Bouten, W., Sorooshian, S., 2003a: Effective and efficient algorithm for multiobjective optimization of hydrologic models. Water Resources Research, 39(8).

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

Wahr, J., Molenaar, M., Bryan, F., 1998: Time variability of the Earth's gravity field: Hydrological and oceanic effects and their possible detection using GRACE. Journal of Geophysical Research: Solid Earth, 103(B12): 30205-30229.

Waltz, R.A., Morales, J.L., Nocedal, J., Orban, D., 2006: An interior algorithm for nonlinear optimization that combines line search and trust region steps. Mathematical programming, 107(3): 391-408.

Watkins, M.M., Wiese, D.N., Yuan, D.N., Boening, C., Landerer, F.W., 2015: Improved methods for observing Earth's time variable mass distribution with GRACE using spherical cap mascons. Journal of Geophysical Research: Solid Earth, 120(4): 2648-2671.

Watkins, M.M., Yuan, D.-N., 2012: JPL Level-2 Processing Standards Document For Level-2 Product Release 05. Jet Propulsion Laboratory, Pasadena, CA, USA.

Webb, R.W., Rosenzweig, C.E., Levine, E.R., 2000: Global Soil Texture and Derived Water-Holding Capacities Dataset. <https://daac.ornl.gov/cgi-bin/dsviewer.pl?ds_id=548>.

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

Węglarczyk, S., 1998: The interdependence and applicability of some statistical quality measures for hydrological models. Journal of Hydrology, 206(1): 98-103.

Werth, S., Güntner, A., Petrovic, S., Schmidt, R., 2009: Integration of GRACE mass variations into a global hydrological model. Earth and Planetary Science Letters, 277(1): 166-173.

Widén-Nilsson, E., Halldin, S., Xu, C.-y., 2007: Global water-balance modelling with WASMOD-M: Parameter estimation and regionalisation. Journal of Hydrology, 340(1): 105-118.

Wielicki, B.A., Barkstrom, B.R., Harrison, E.F., Lee III, R.B., Louis Smith, G., Cooper, J.E., 1996: Clouds and the Earth's Radiant Energy System (CERES): An earth observing system experiment. Bulletin of the American Meteorological Society, 77(5): 853-868.

Wiese, D.N., Yuan, D.-N., Boening, C., Landerer, F.W., Watkins, M.M., 2015: GRACE Mascon Ocean, Ice, and Hydrology Equivalent Water Height JPL RL05M.1. Ver. 1 Dataset. <http://dx.doi.org/10.5067/TEMSC-OCL05>.

Winker, P., Gilli, M., 2004: Applications of optimization heuristics to estimation and modelling problems. Computational Statistics & Data Analysis, 47(2): 211-223.

Wisser, D., Fekete, B., Vörösmarty, C., Schumann, A., 2010: Reconstructing 20th century global hydrography: a contribution to the Global Terrestrial Network-Hydrology (GTN-H). Hydrology and Earth System Sciences, 14(1): 1-24.

Wood, E.F., Roundy, J.K., Troy, T.J., Van Beek, L., Bierkens, M.F., Blyth, E., de Roo, A., Döll, P., Ek, M., Famiglietti, J., 2011: Hyperresolution global land surface modeling: Meeting a grand challenge for monitoring Earth's terrestrial water. Water Resources Research, 47(5).

Xie, H., Longuevergne, L., Ringler, C., Scanlon, B., 2012: Calibration and evaluation of a semi-distributed watershed model of Sub-Saharan Africa using GRACE data. Hydrology and Earth System Sciences, 16(9): 3083-3099.

Xu, C., 2002: Hydrologic Models. Department of Earth Sciences Hydrology, Uppsala University Uppsala.

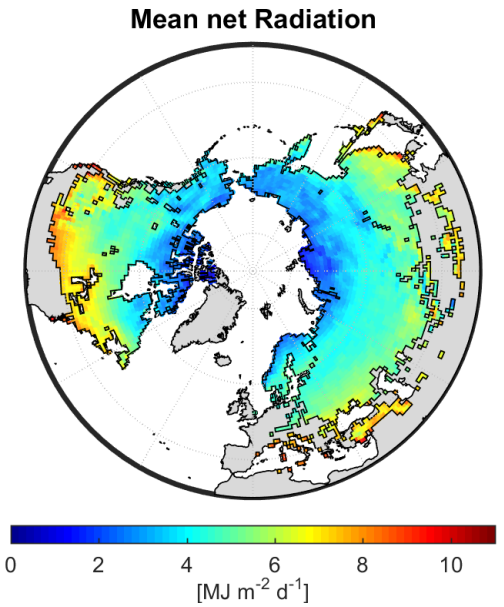
Yapo, P.O., Gupta, H.V., Sorooshian, S., 1998: Multi-objective global optimization for hydrologic models. Journal of Hydrology, 204(1): 83-97.

Zeng, N., Yoon, J.-H., Mariotti, A., Swenson, S., 2008: Variability of basin-scale terrestrial water storage from a PER water budget method: The Amazon and the Mississippi. Journal of Climate, 21(2): 248-265.

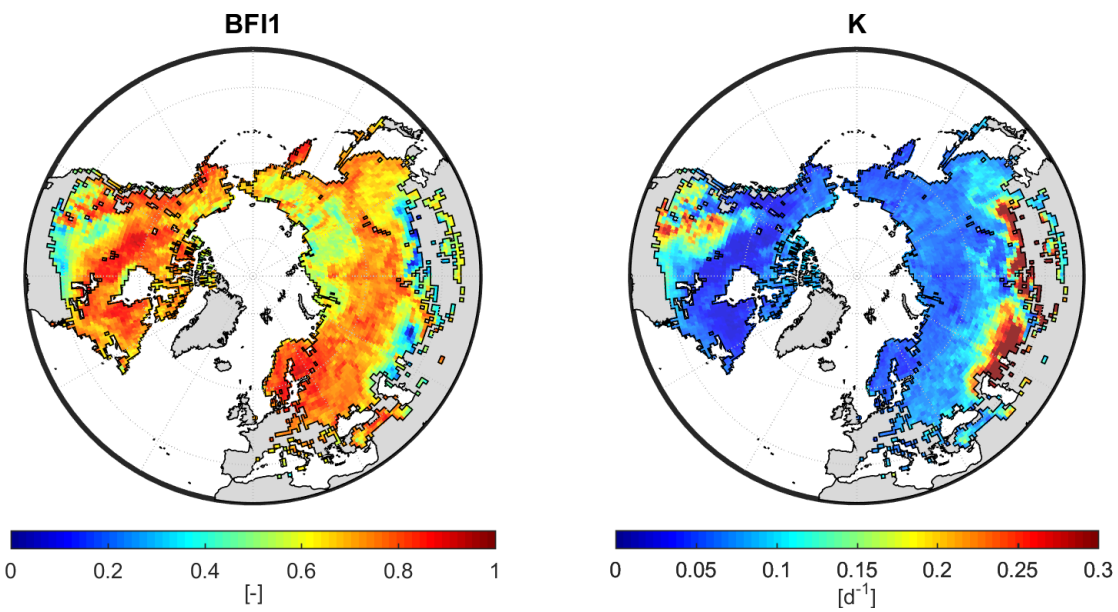
Zhang, L., Potter, N., Hickel, K., Zhang, Y., Shao, Q., 2008: Water balance modeling over variable time scales based on the Budyko framework: Model development and testing. Journal of Hydrology, 360(1): 117-131.

Appendix

I) Study Area



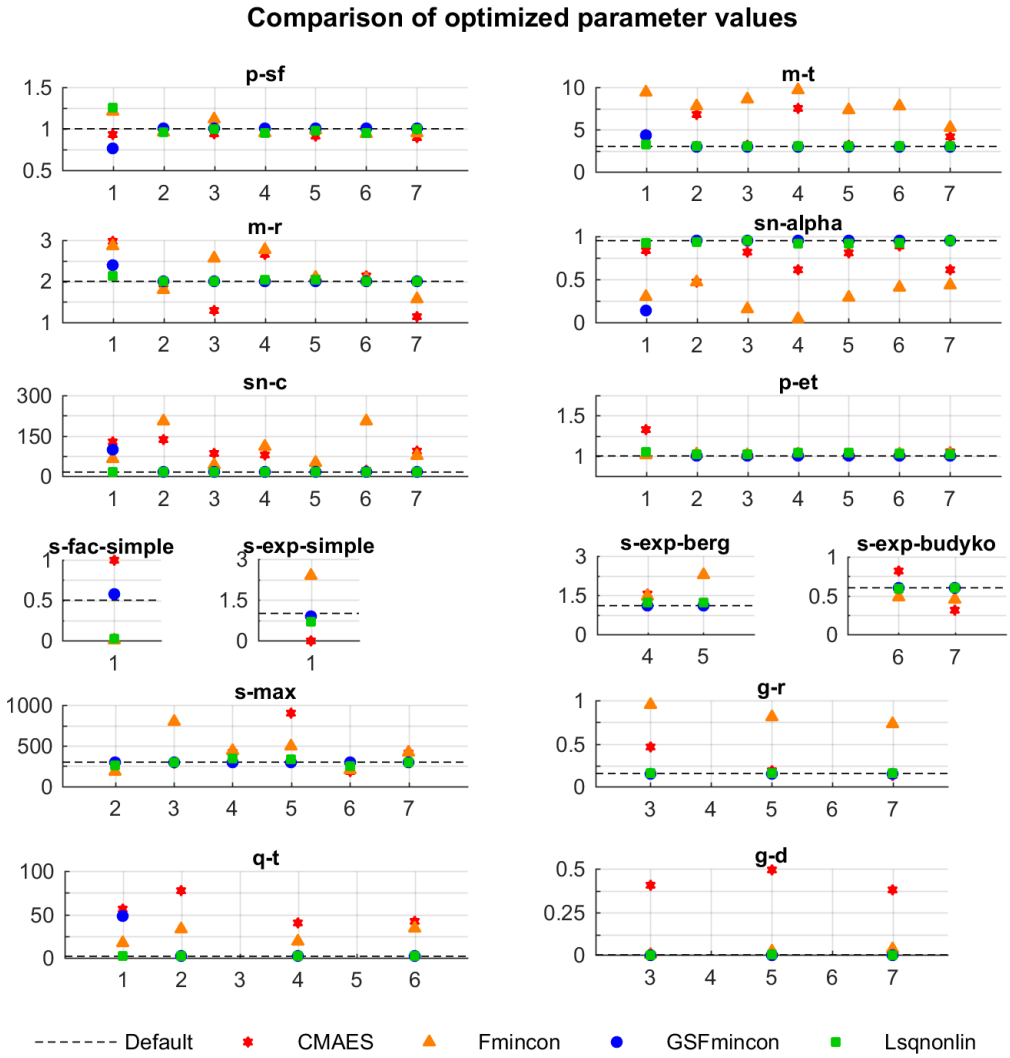
A 1: Average net radiation [MJ m^{-2} d^{-1}] of the study area obtained from the meteorological forcing data for the period 2003 – 2010.



A 2: Base flow Index 1 (BFI1) and base flow recession constant K [d^{-1}] for the study area as provided by Beck et al. (2015).

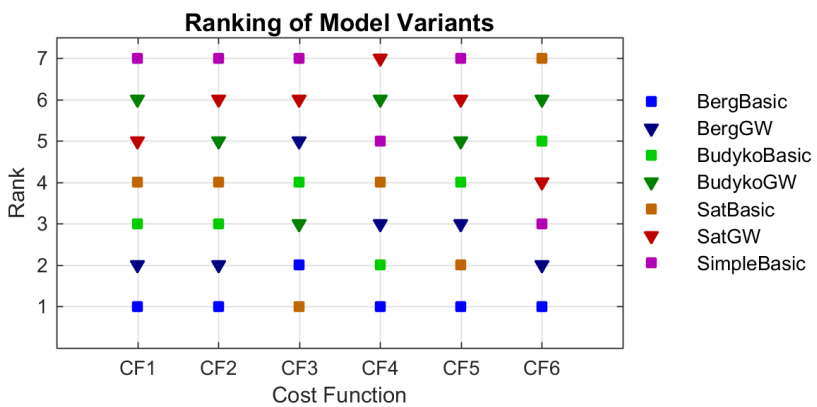
II) Comparison of Optimization Methods

Optimized Parameter Values of tested Search Algorithm



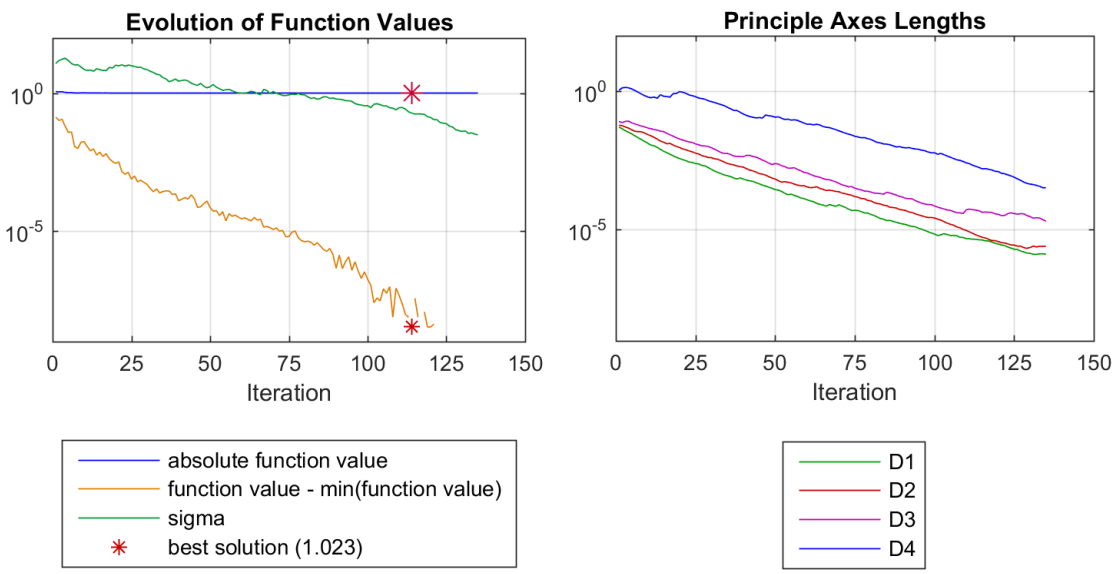
A 3: Comparison of optimized parameter values applying Lsqnonlin, Fmincon, GSFmincon and CMAES with CF1 for 1000 grid cells during testing of optimization methods.

Model Ranking according to Total Costs for each Cost Function

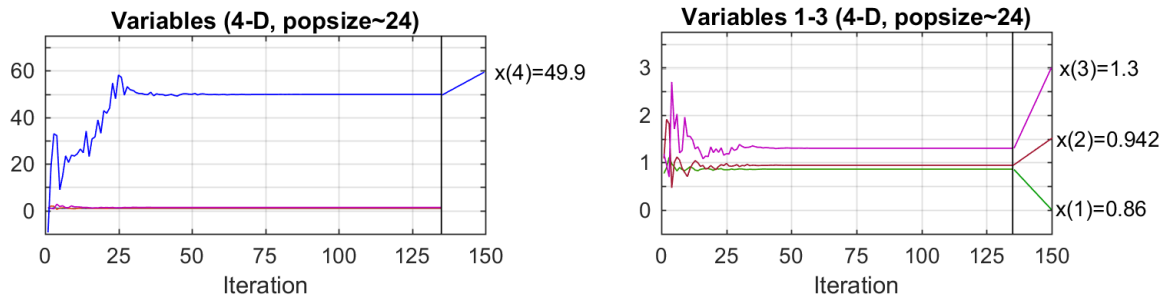


A 4: Ranking of model variants according to total costs obtained with different cost functions (CF1-CF6) during testing of cost functions (applying CMAES for 100 grid cells).

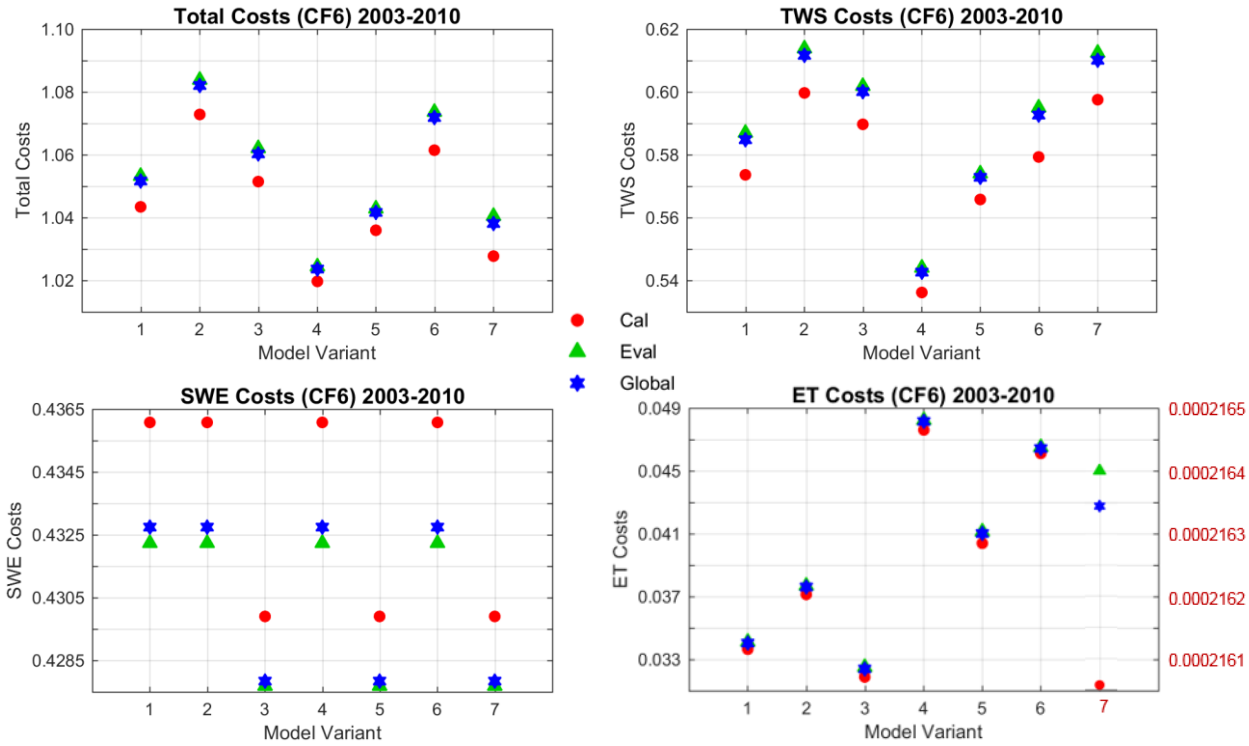
III) Final Model Calibration



A 5: Characteristics of the final calibration of BergBasic. (CMAES, CF6, 1000 grid cells) left: Evolution of function values (total costs), the difference of the current function value to the overall minimum found, and step size sigma as well as the final best solution. The graph shows that the function value decreases with each iteration, while the improvement is considerably low (< 10^{-5}) and thus the optimization process stopped after approximately 120 iterations. right: Evolution of the length of principle axes of CMAES' distribution ellipsoid for each (parameter) dimension. Reduction of all axes length indicates that the population converges to the minimum.

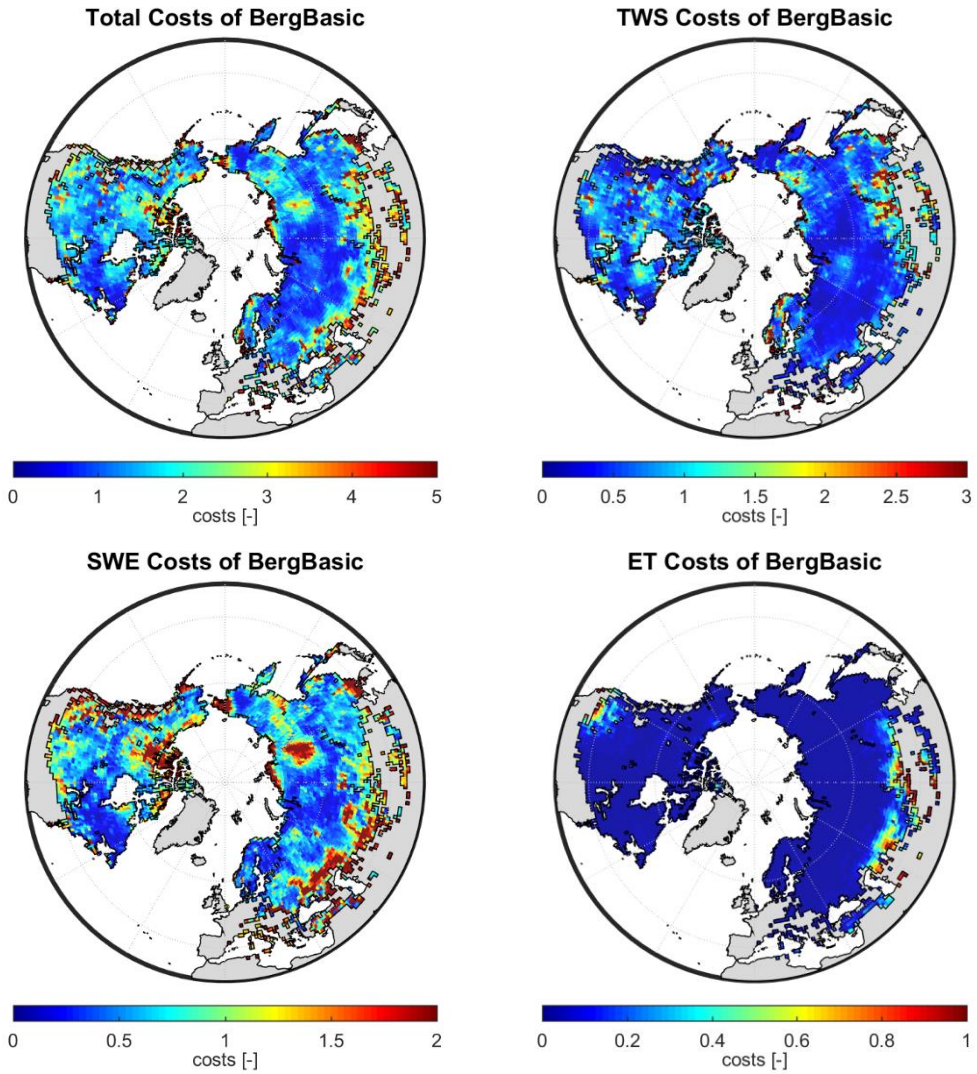


A 6: Stabilization of parameter values during optimization (BergBasic). x indicate the optimized scaling factors associated with (1) p_et, (2) s_max, (3) s_exp_berg and (4) q_t.

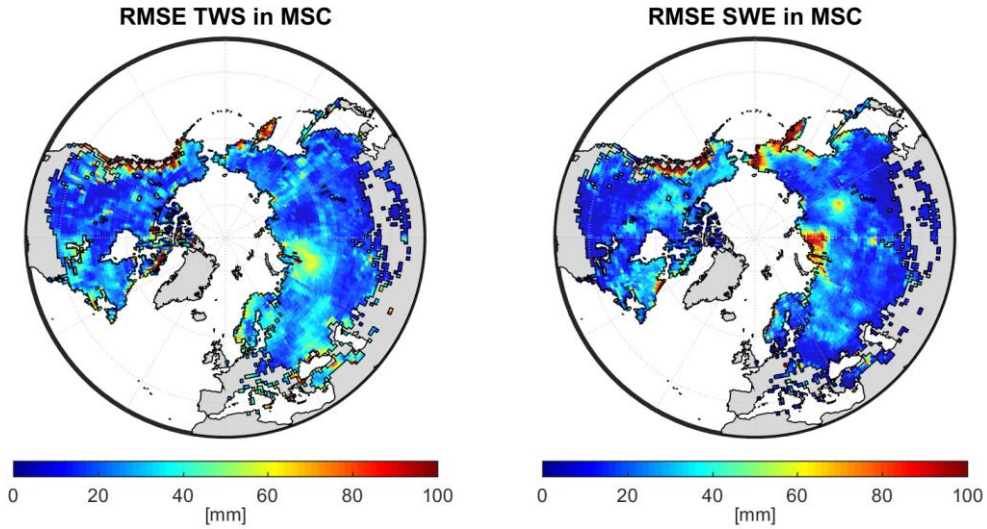


A 7: Total, TWS, SWE and ET costs for calibration (Cal), evaluation (Eval), and the entire study area (Global) for the period 2003 – 2010. Model variants: 1-SimpleBasic, 2-SatBasic, 3-SatGW, 4-BergBasic, 5-BergGW, 6-BudykoBasic, 7-BudykoGW; optimization for 1000 grid cells using CMAES and CF1.

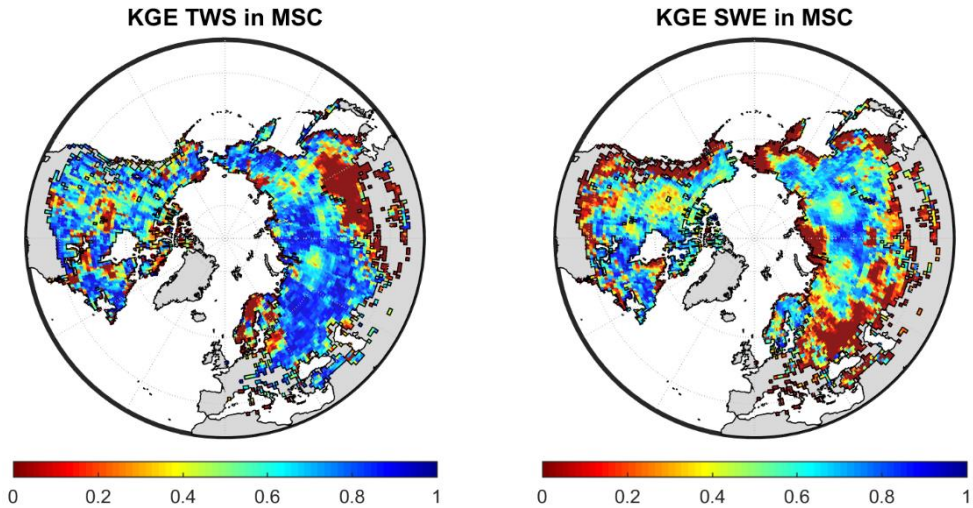
IV) Pixelwise Costs 2004 – 2010 of BergBasic



A 8: Total, TWS, SWE and ET costs (CF6) of optimized BergBasic model for each grid cell for the period 2004 – 2010.

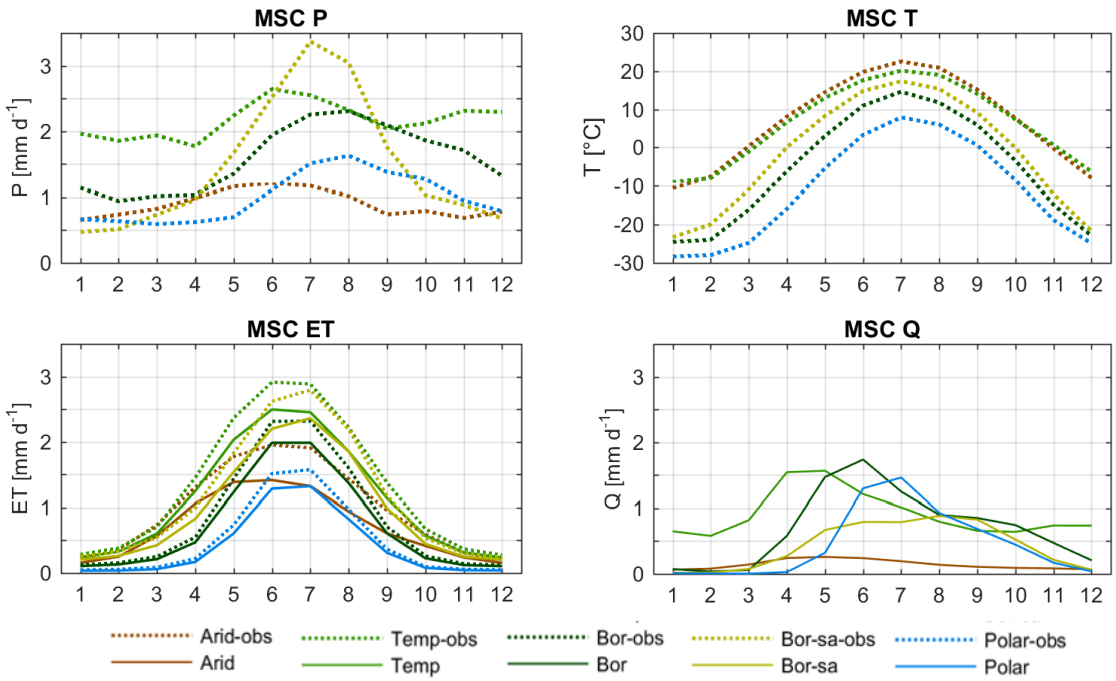


A 9: RMSE of mean seasonal cycle between TWS_{obs} and TWS_{mod} (left) an) SWE_{obs} and SWE_{mod} (right) for each grid cell for the period 2004 – 2010.

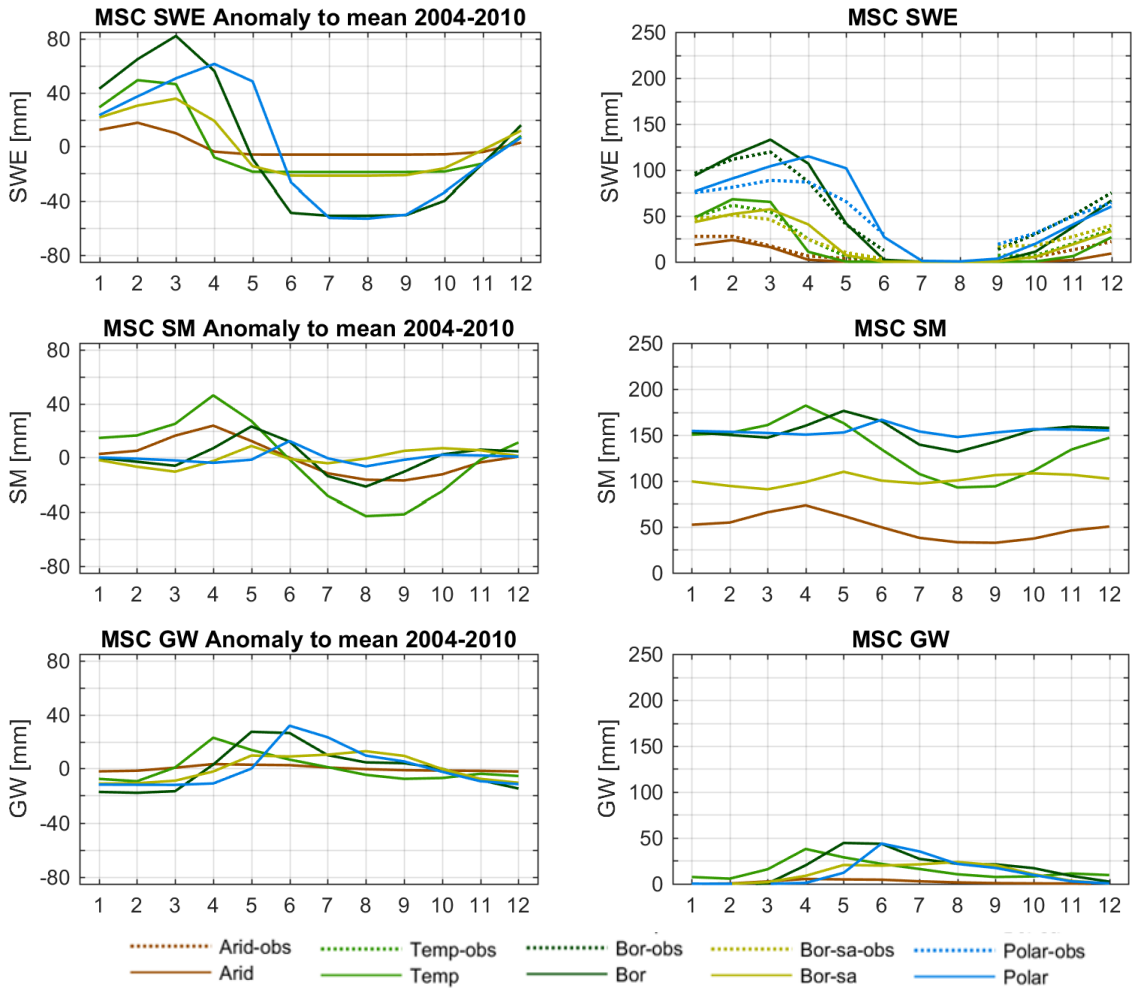


A 10: KGE of mean seasonal cycle between TWS_{obs} and TWS_{mod} (left) and SWE_{obs} and SWE_{mod} (right) for each grid cell for the period 2004 – 2010.

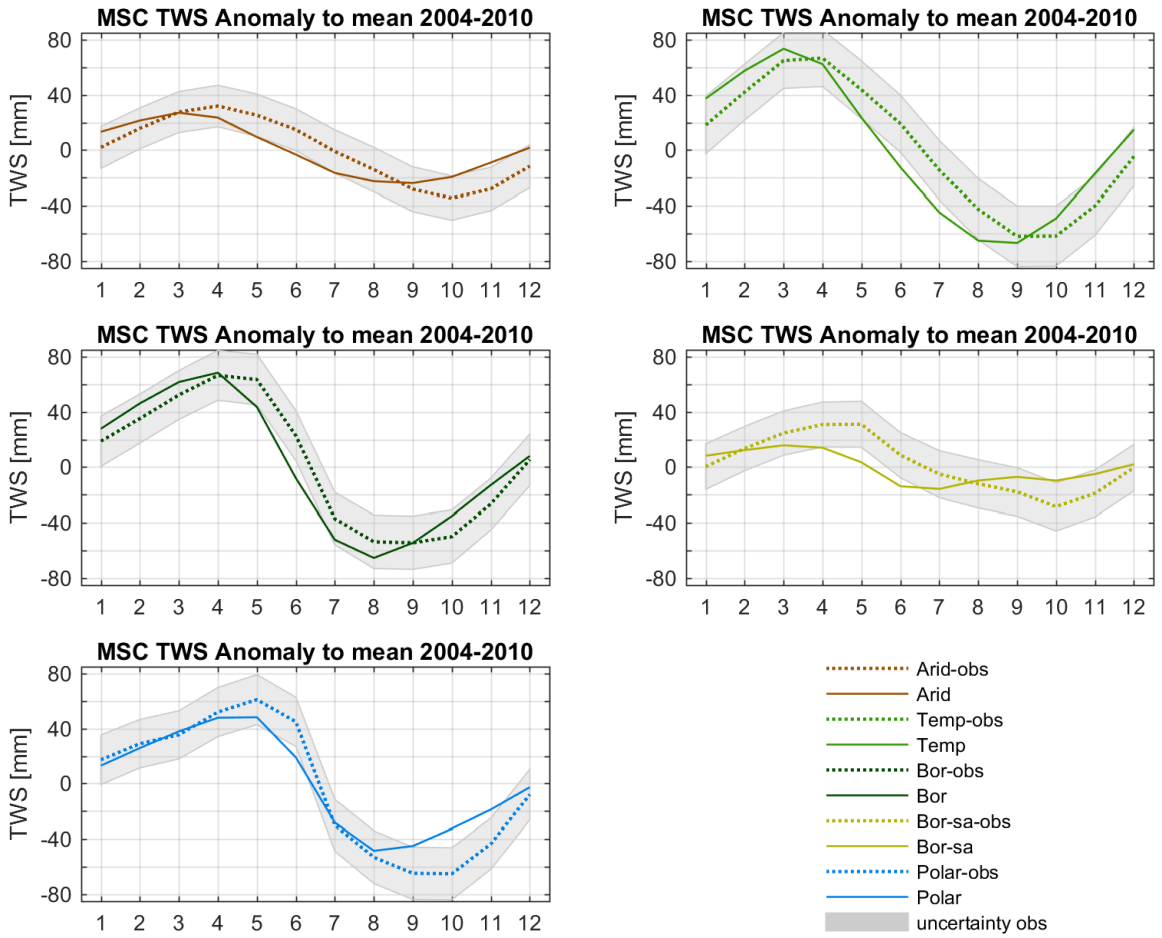
V) Mean Seasonal Cycle of KG zones



A 11: Comparison of the mean seasonal cycle (MSC) of precipitation P [mm d^{-1}], temperature T [°C], evapotranspiration ET [mm d^{-1}] and runoff Q [mm d^{-1}] for each KG zone. Dotted lines: observations; solid lines: simulations.

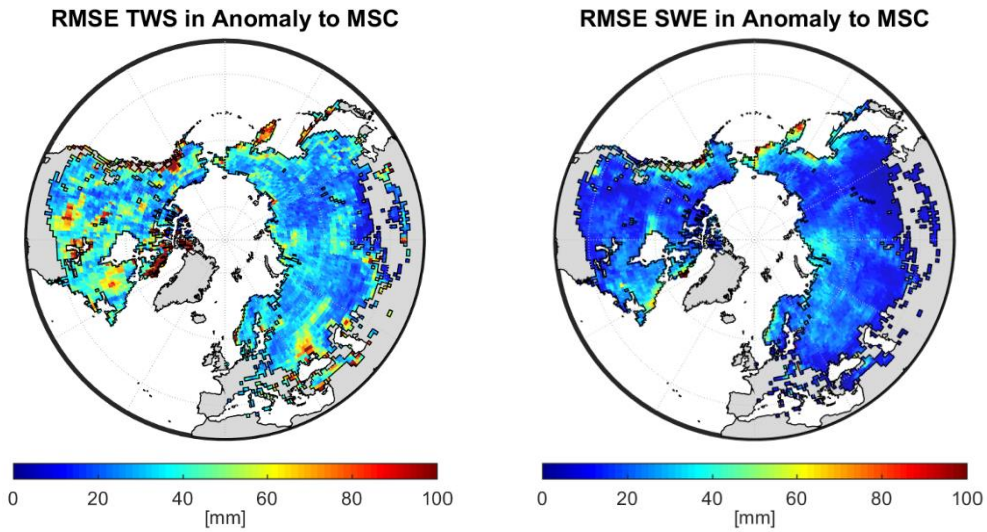


A 12: Comparison of the mean seasonal cycle of snow water equivalent SWE [mm], soil moisture SM [mm] and groundwater GW [mm] (right) as well as the mean seasonal cycle of the corresponding storage anomalies to the baseline 2004 – 2010 for each KG zone. Dotted lines: observations; solid lines: simulations.

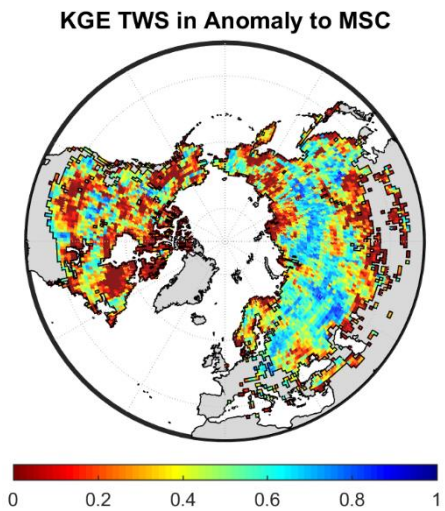


A 13: Comparison of the mean seasonal cycle of observed and simulated terrestrial water storage anomalies TWS_{obs} (dotted lines) and TWS_{mod} (solid lines) [mm] to the baseline 2004 – 2010 for each KG zone. The uncertainty estimates of TWS_{obs} are depicted as grey shaded area.

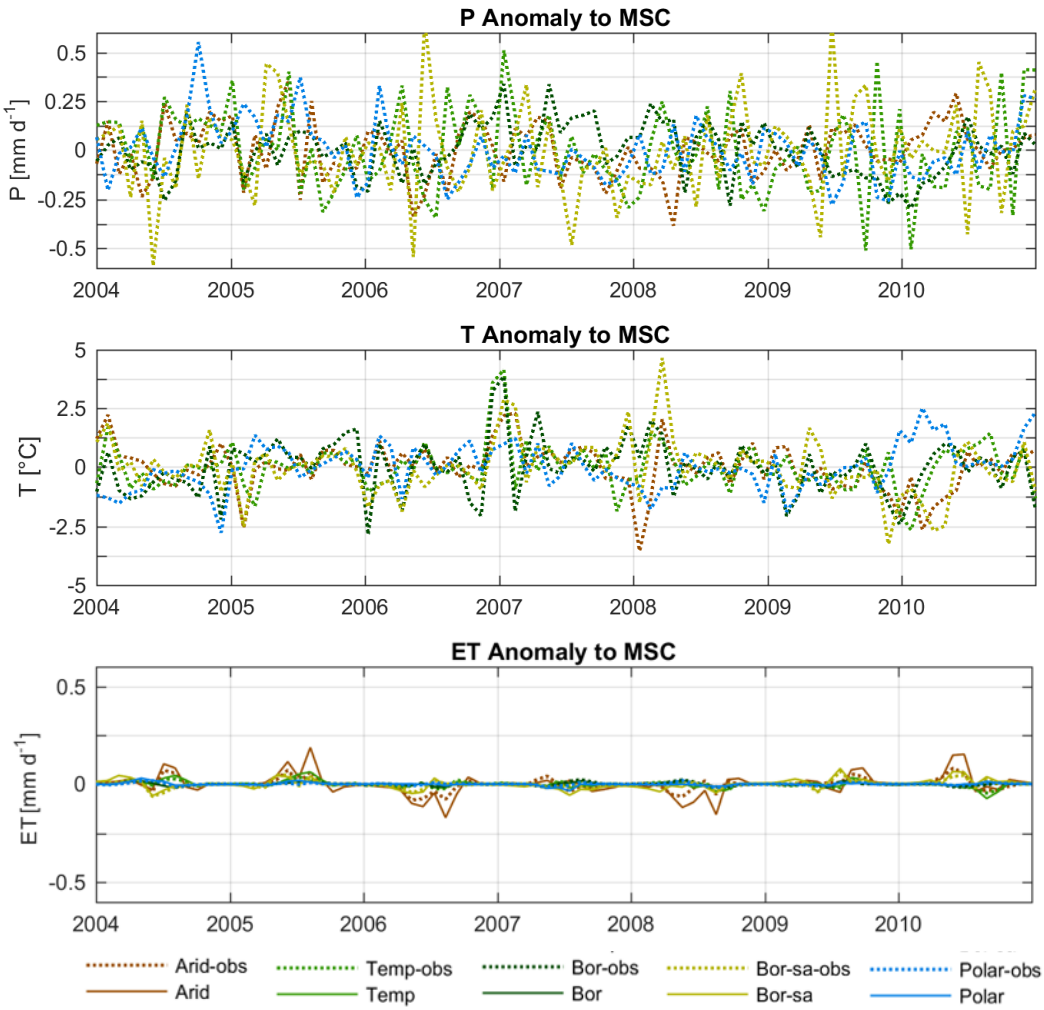
VI) Anomaly to Mean Seasonal Cycle of KG zones



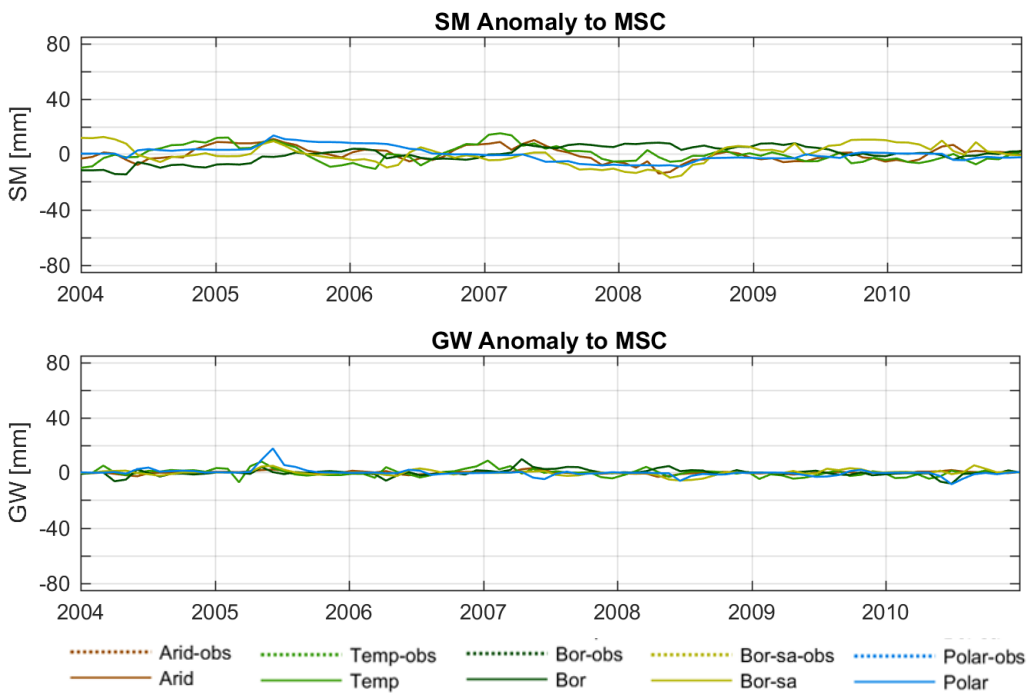
A 14: RMSE of anomaly to mean seasonal cycle between TWS_{obs} and TWS_{mod} (left) and SWE_{obs} and SWE_{mod} (right) for each grid cell for the period 2004 – 2010.



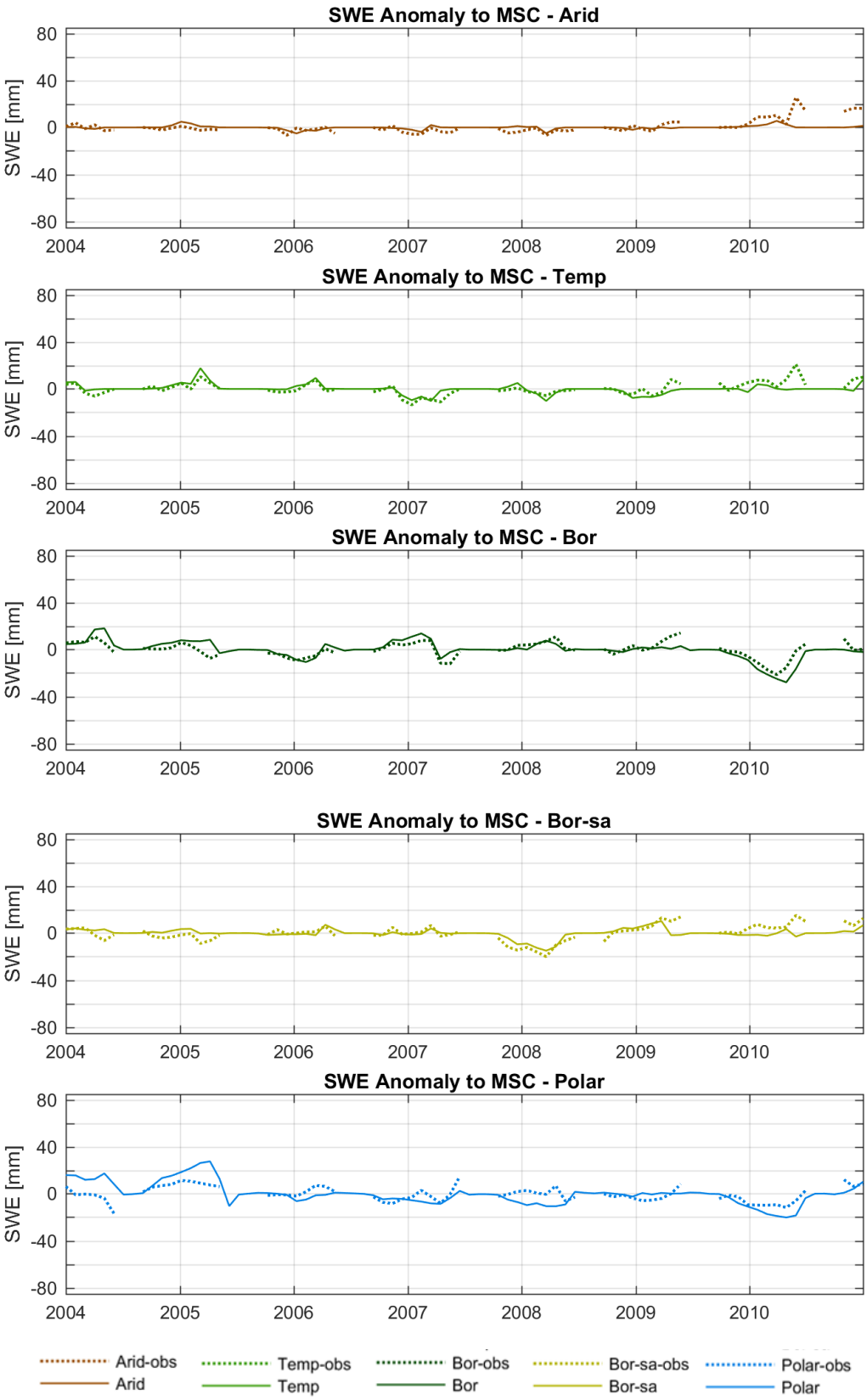
A 15: KGE of anomaly to mean seasonal cycle between TWS_{obs} and TWS_{mod} for each grid cell for the period 2004 – 2010.



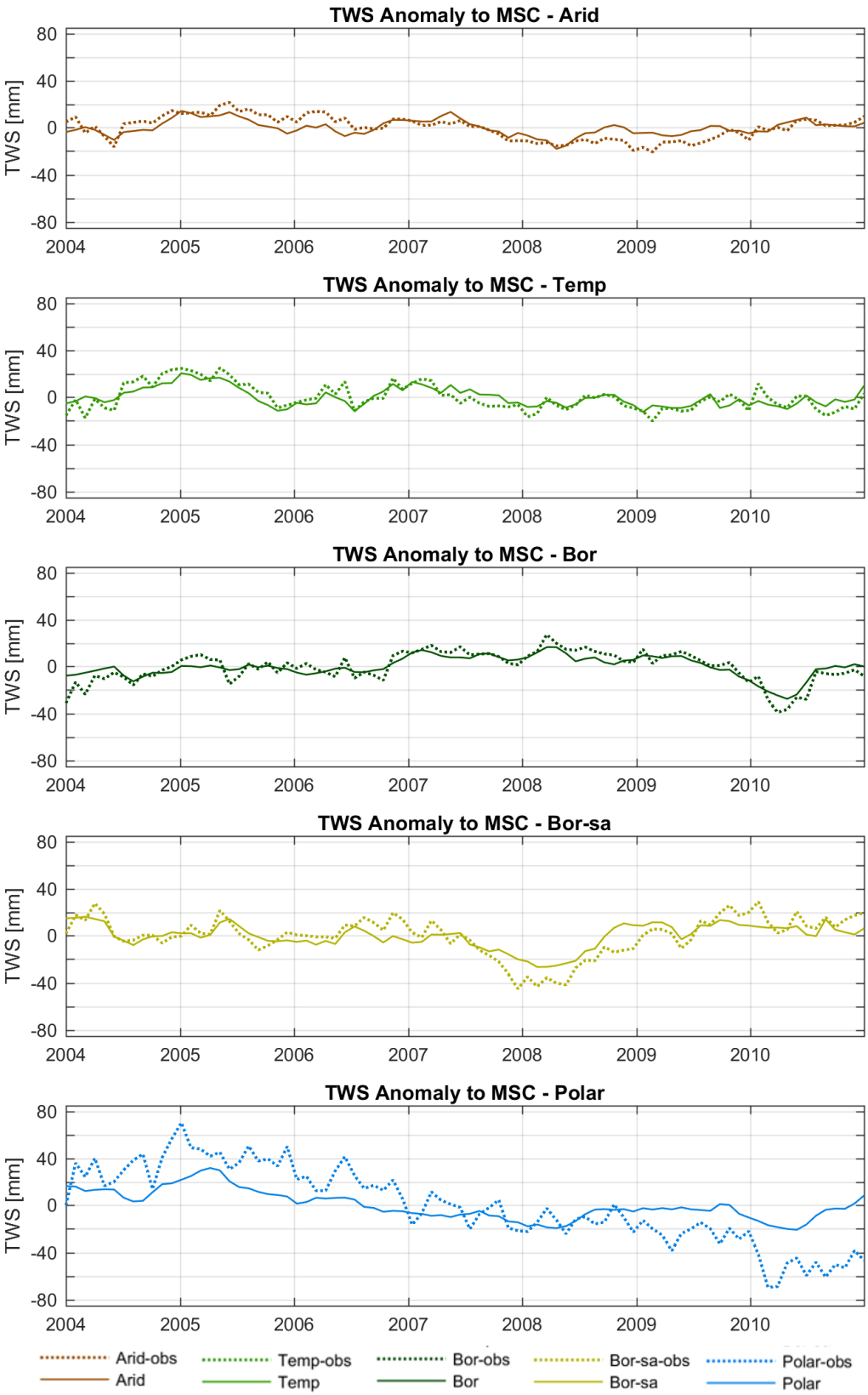
A 16: Anomaly of precipitation P [mm d^{-1}], temperature T [°C], and evapotranspiration ET [mm d^{-1}] to the mean seasonal cycle 2004 – 2010 for each KG zone. Dotted lines: observations; solid lines: simulations.



A 17: Anomaly of SM and GW to the mean seasonal cycle 2004 – 2010 for each KG zone.



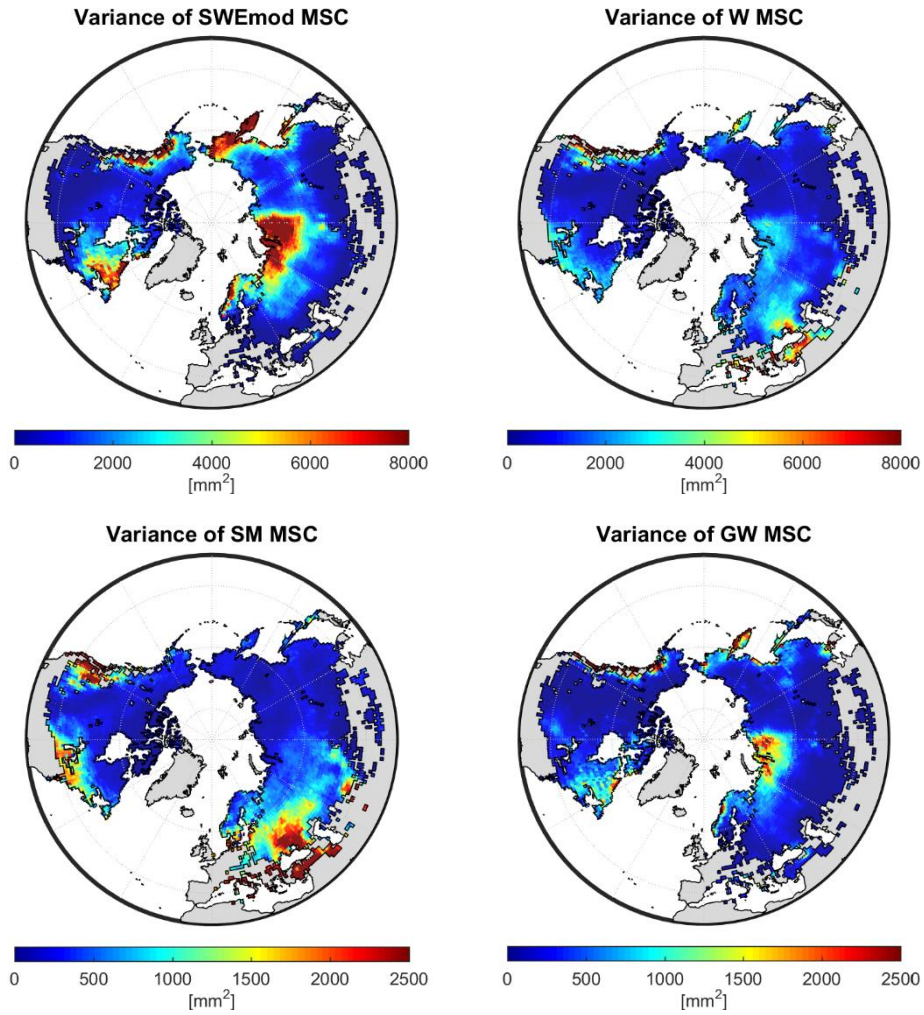
A 18: Anomaly of SWE_{obs} (dotted lines) and SWE_{mod} (solid lines) [mm] to the mean seasonal cycle 2004 – 2010 for each KG zone.



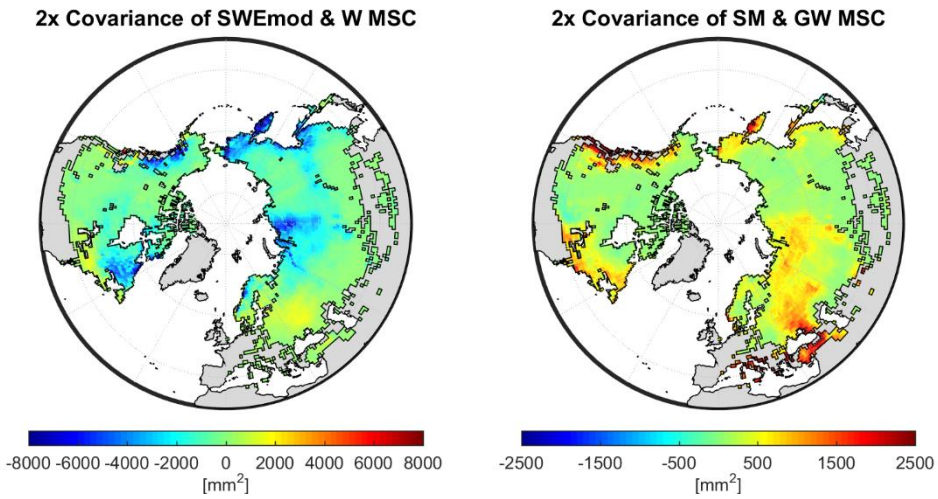
A 19: Anomaly of TWS [mm] to the mean seasonal cycle 2004 – 2010 for each KG zone.

VII) Composition of TWS

Variances in Mean Seasonal Cycle

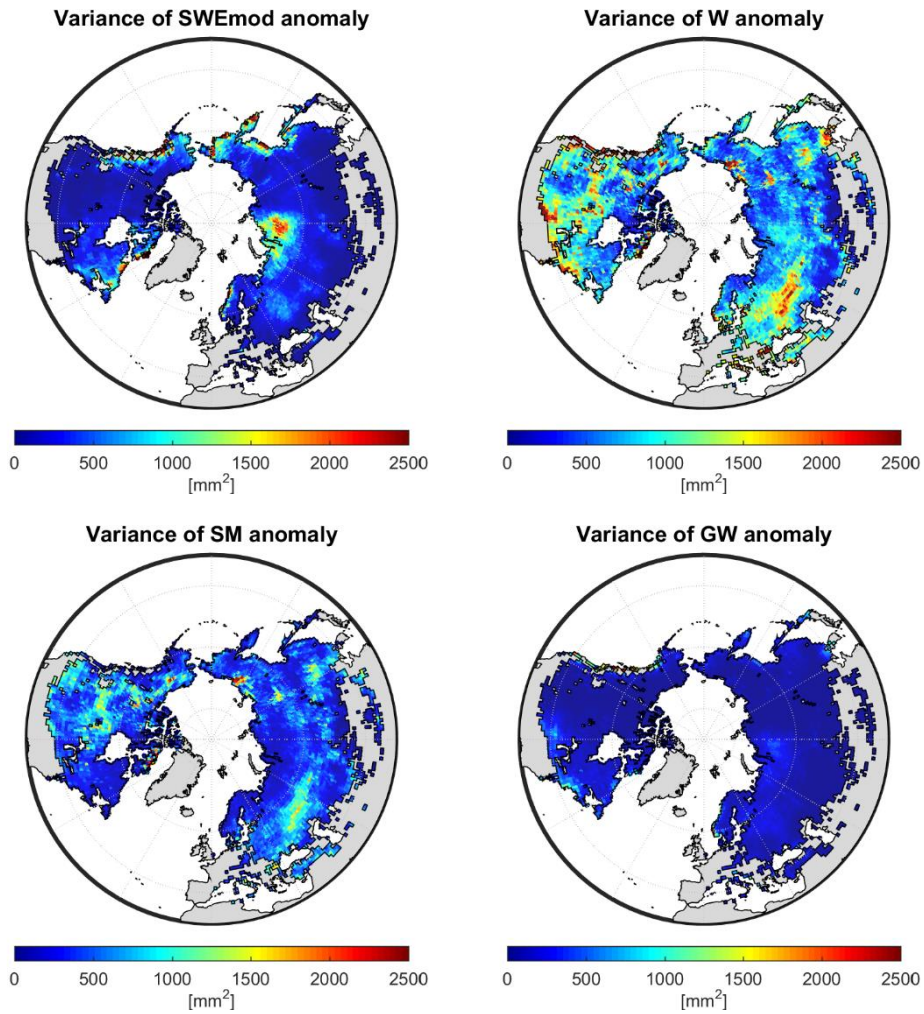


A 20: Variance in the mean seasonal cycle of simulated snow water equivalent SWE_{mod}, total available water W (SM + GW), soil moisture SM and groundwater GW of each grid cell for the period 2004 – 2010.

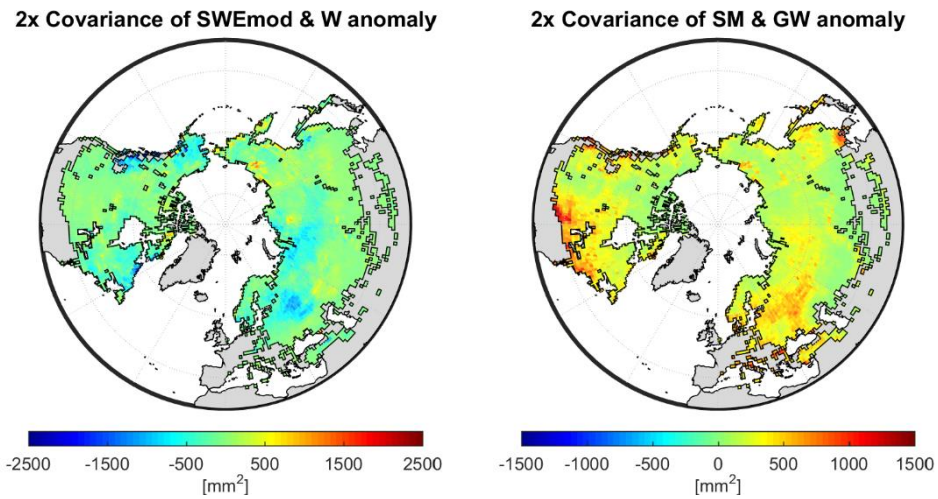


A 21: Two times the covariance between the mean seasonal cycle of SWE_{mod} and W resp. SM and GW of each grid cell for the period 2004 – 2010. Positive values/red colors indicate a positive correlation, while negative values/blue colors indicate a negative correlation.

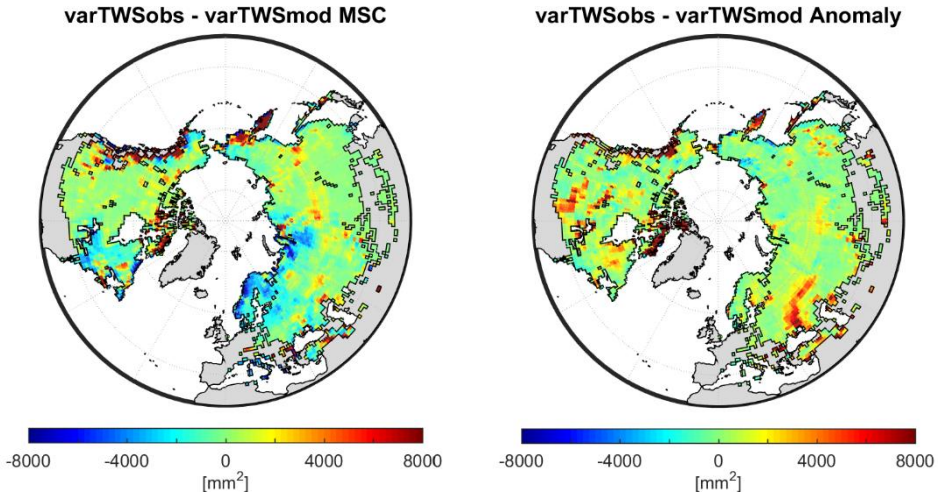
Variances in Anomalies to Mean Seasonal Cycle



A 22: Variance in the anomaly to the mean seasonal cycle of simulated snow water equivalent SWE_{mod}, total available water W (SM + GW), soil moisture SM and groundwater GW of each grid cell for the period 2004 – 2010.

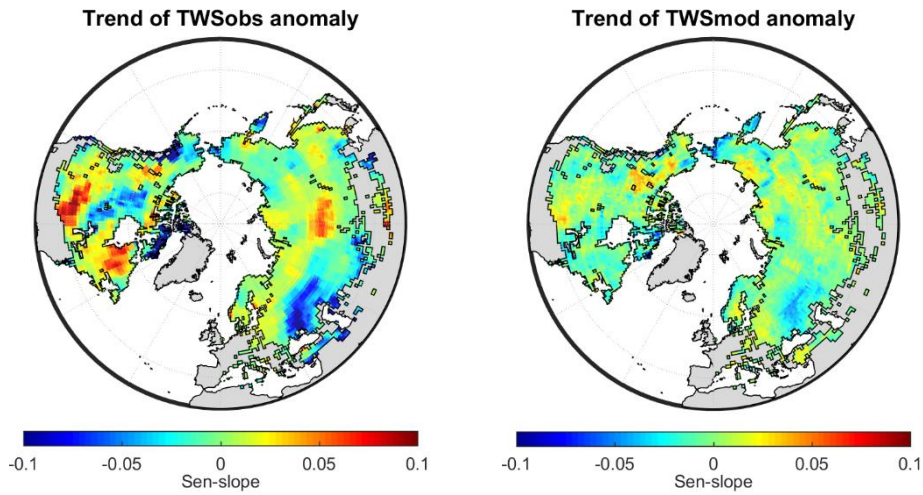


A 23: Two times the covariance between the anomaly to mean seasonal cycle of SWE_{mod} and W resp. SM and GW of each grid cell for the period 2004 – 2010. Positive values/red colors indicate a positive correlation, while negative values/blue colors indicate a negative correlation.

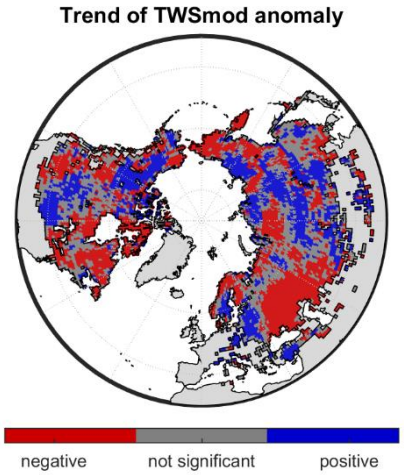


A 24: Differences between the variances $varTWS_{obs}$ and $varTWS_{mod}$ of the mean seasonal cycle (left) and anomalies to the mean seasonal cycle (right) for each grid cell and the period 2004 – 2010. Negative values/blue colors indicate that the model overestimates TWS variability, while positive values/red color suggest underestimation of TWS_{obs} variance by the model.

VIII) Trends in TWS Anomaly



A 25: Trend (Sen-slope) in TWS_{obs} (left) and TWS_{mod} (right) anomalies to the mean seasonal cycle (right) for each grid cell and the period 2004 – 2010. Negative values/blue colors indicate a negative trend, while positive values/red color suggest a positive trend. Significance is not considered.



A 26: Qualitative trends in the anomaly to the mean seasonal cycle of TWS_{mod} for the period 2004 – 2010.

A 27: Confusion matrix of randomly generated trend classes and qualitative trends in TWS_{obs} anomaly. With three possible classes, the overall accuracy by chance is 33 %.

		obs			Total	User Acc
		neg	not sig	pos		
random	neg	754	787	750	2291	0.50
	not sig	488	479	520	1487	0.37
	pos	837	782	793	2412	0.38
Total		2333	2079	2048	2063	6190
Producer Acc		0.49	0.36	0.23	0.38	**0.33**

Selbstständigkeitserklärung

Ich versichere hiermit, dass ich die vorliegende Masterarbeit mit dem Titel

„Macroscopic diagnostic modeling of the hydrological cycle:
Understanding the dynamics of water pools in snow affected regions.“

eigenständig und nur unter Verwendung der angegebenen Hilfsmittel und Quellen angefertigt habe. Die eingereichte Arbeit ist nicht anderweitig als Prüfungsleistung verwendet worden oder in deutscher oder einer anderen Sprache als Veröffentlichung erschienen.

Jena, den 15.07.2016

_____

Tina Trautmann