University of Stuttgart
Institute for Modelling Hydraulic and Environmental Systems
Department of Hydrology and Geohydrology
Prof. Dr. rer. nat. Dr.-Ing. András Bárdossy

Master Thesis

Evaluation of precipitation products in Norway

Submitted by

Jessica Sienel

24. May 2022

In cooperation with:

SINTEF

Supervisor: M. Sc. Lennart Schönfelder, Dr. rer.nat. Jochen Seidel

Examiner: Prof. Dr. rer.nat. Dr.-Ing. András Bárdossy

Declaration

I hereby certify that I have written this master thesis independently and not have used sources and aids other than those specified. All statements taken literally or analogously from other writings are marked as such. Other people were not involved in the production of the content of this work. So far, the work has not been published in the same or a similar form, either at home or abroad submitted to another testing authority.

_____ _____

Place, Date Signature

Abstract

Gathering accurate precipitation data is an important task for setting up hydrological models. In Norway, the gauge network density is higher in the southern parts and decreases in the north. The amount of precipitation gauges in higher elevations is rather scarce. Radar data is available, but lacks accurate reflectivity-precipitation relation and errors in precipitation estimation are caused for example by beam blockage.

For modelling purposes, this study aims to evaluate the impact of precipitation data and whether the application of radar derived data gives any benefit, especially when modelling in hourly timesteps. The results of this study can give decision support for modellers having difficulties choosing between precipitation products. For that purpose, spatial interpolated precipitation datasets were evaluated and compared in terms of performance for hydrological modelling. The Meteorological Institute Norway publishes gridded hourly datasets covering the Norwegian mainland: They include seNorge2, interpolated gauge data using an optimal interpolation, and the MEPS-Nordic (NWP), a combination of gauge data, radar data and a numerical weather model. Five different catchments were simulated in the numerical precipitation-runoff model HYPE with both datasets for comparison. The catchments vary in area, hydrological regime and availability of nearby gauges; they are mostly unregulated. The simulation was done in an hourly time step in order to compare precipitation variability in detail.

In this study, a calibration method was developed that generates comparable and reproduceable performances in terms of the Kling–Gupta efficiency (KGE) for each catchment and dataset. A Progressive Monte Carlo Method within a stepwise calibration was applied. The resulting discharge time series and water balances of the catchments were analysed and compared. Additionally, selected precipitation events, where the precipitation products were not able to describe atmospheric processes appropriately, were analysed in-depth. The datasets were further compared by spatially accumulating annual precipitation sums over the catchments, by using a weather station owned by a hydro power company to evaluate the agreement of the data and by comparing the runoff and precipitation volume of the basins.

It turned out that applying raw radar data is not reasonable because of beam shielding and attenuation effects, especially in the inland of Norway, as well as a high underestimation of the precipitation amount. SeNorge2 showed reasonable results in the calibration, but decreased strongly in validation. NWP appeared to create more stable model performances and had a better agreement in terms of water balance and detection of small-scale precipitation events. Nevertheless, the model performance is lower in most of the catchments.

During a sensitivity analysis, different model uncertainties were found and discussed. They included unsensitive parameters, that did not converge during the calibration. Additionally, switching the calibration and validation time period demonstrated instabilities of the model results depending on the chosen time period in some catchments. Individual parameters were evaluated as well. Some of them showed physical reasonable values, but for example the river velocity appeared to be unnaturally high in some instances.

NWP turned out to be more likely to measure small scale precipitation events and resulted in more stable performances and was therefore more suitable for an hourly hydrological model. However, the amount radar stations in Norway should be increased, to decrease attenuation and beam blockage effects.

Kurzfassung

Zu Modellierungszwecken zielt diese Studie darauf ab, die Auswirkungen von Niederschlagsdaten zu bewerten und festzustellen, ob die Anwendung von Radardaten einen Nutzen bringt, insbesondere bei der Modellierung in stündlichen Zeitschritten. Die Ergebnisse dieser Studie können Modellierern, die Schwierigkeiten bei der Auswahl zwischen Niederschlagsprodukten haben, eine Entscheidungshilfe geben. Zu diesem Zweck wurden räumlich interpolierte Niederschlagsdatensätze aus Norwegen ausgewertet und hinsichtlich ihrer Leistungsfähigkeit für die hydrologische Modellierung verglichen.

In Norwegen ist das Niederschlagsmessnetz im Süden dichter und nimmt im Norden ab. Die Anzahl an Niederschlagsmessern im Hochgebirge ist spärlich. Radardaten sind verfügbar, aber es mangelt an einer genauen Reflektivitäts-Niederschlagsbeziehung und Fehler bei der Niederschlagsmessung werden beispielsweise durch Strahlblockierung verursacht. Das Meteorologische Institut Norwegen veröffentlicht Norwegen umfassende Niederschlagsprodukte. Dazu gehört seNorge2, eine Interpolation von Stationsdaten, und MEPS-Nordic (NWP), eine Kombination aus Stationsdaten, Radardaten und einem numerischen Wettermodell. Im numerischen Niederschlagsabflussmodell HYPE wurden fünf verschiedene Einzugsgebiete mit diesen Datensätzen simuliert und die Ergebnisse verglichen. Die Einzugsgebiete unterscheiden sich in Fläche, hydrologischem Regime und Verfügbarkeit nahegelegener Messstationen und sind meist ungeregelt. Die Simulation wurde in einem stündlichen Zeitschritt durchgeführt, um die Niederschlagsvariabilität im Detail vergleichen zu können.

In dieser Studie wurde eine Kalibrierungsmethode entwickelt, die eine vergleichbare und reproduzierbare Modellperformance in Bezug auf die Kling-Gupta-Effizienz (KGE) für jedes Einzugsgebiet und jeden Datensatz generiert. Es wurde eine progressive Monte-Carlo-Methode innerhalb einer schrittweisen Kalibrierung angewendet. Die resultierenden Abflusszeitreihen und Wasserbilanzen der Einzugsgebiete wurden analysiert und verglichen. Zusätzlich wurden ausgewählte Niederschlagsereignisse, bei denen die Niederschlagsprodukte atmosphärische Prozesse nicht angemessen beschreiben konnten, vertieft analysiert. Die Datensätze wurden darüber hinaus verglichen, indem die jährlichen Niederschlagssummen räumlich über den Einzugsgebieten akkumuliert, eine Wetterstation mit den Produkten verglichen und die Wasserbilanz der Einzugsgebiete aufgestellt wurden.

Es stellte sich heraus, dass die Verwendung von Radardaten aufgrund von Strahlblockierung- und Dämpfungseffekten insbesondere im inneren Norwegens sowie einer starken Unterschätzung der Niederschlagsmenge nicht sinnvoll ist. Ein Modell mit seNorge2 zeigte bei der Kalibrierung passable Ergebnisse, ließ aber bei der Validierung stark nach. NWP schien stabilere Modellperformances zu erzeugen und hatte eine bessere Übereinstimmung in Bezug auf die Wasserbilanz und die Erkennung von kleinräumigen Niederschlagsereignissen. Dennoch ist die Modellperformance in den meisten Einzugsgebieten geringer als bei seNorge2.

Bei einer Sensitivitätsanalyse wurden verschiedene Modellunsicherheiten gefunden und diskutiert. Darunter Parameter, die während der Kalibrierung nicht konvergierten und nicht sensitiv sind. Darüber hinaus zeigte das Wechseln des Kalibrierungs- und Validierungszeitraums Instabilitäten der Modellergebnisse in Abhängigkeit vom gewählten Zeitraum in einigen Einzugsgebieten. Auch einzelne Parameter wurden ausgewertet. Einige von ihnen zeigten physikalisch vernünftige Werte, andere Parameter, wie die Flussgeschwindigkeit, waren hingegen unnatürlich hoch.

Es stellte sich heraus, dass NWP für stündliche Modelle besser geeignet ist, da mehr Niederschlagsereignisse tatsächlich gemessen wurden und die Ergebnisse stabiler sind. Allerdings sollte die Anzahl an Radarstationen in Norwegen erhöht werden, um Dämpfungseffekte und Strahlblockungen zu verringern.

Abstrakt

Å samle nedbørsdata er en viktig oppgave for å sette opp hydrologiske modeller. Nøyaktig informasjon reduserer modelleringsusikkerhet. I Norge er nedbørmålingsnettverket tettere i sør og avtar i nord. Antall regnmålere i høyfjellet er sparsomt. Radardata er tilgjengelig, men mangler nøyaktig reflektivitet-nedbør-relasjon og feil i nedbørestimering skyldes for eksempel stråleblokkering.

For modelleringsformål har denne studien som mål å evaluere virkningen av nedbørsdata og om bruken av radaravledede data gir noen fordel, spesielt ved modellering i timestrinn. Resultatene av denne studien kan gi beslutningsstøtte for modellerere som har vanskeligheter med å velge mellom nedbørsprodukter. For det formålet ble romlig interpolerte nedbørsdatasett evaluert og sammenlignet med tanke på ytelse for hydrologisk modellering. Meteorologisk institutt Norge publiserer rutenettede timedatasett som dekker det norske fastlandet: De inkluderer seNorge2, interpolerte værstasjonsdata, og MEPS-Nordic (NWP), en kombinasjon av værstasjonsdata, radardata og en numerisk værmodell. Fem forskjellige nedslagsfelt ble simulert i den numeriske nedbør-avrenningsmodellen HYPE med begge datasettene for sammenligning. Nedslagsfeltene varierer i område, hydrologisk regime og tilgjengelighet av nærliggende værstasjonsdata; de er stort sett uregulerte. Simuleringen ble gjort i et timestrinn for å sammenligne nedbørvariabilitet i detalj.

I denne studien ble det utviklet en kalibreringsmetode som skaper sammenlignbare og reproduserbare resultater når det gjelder Kling–Gupta-effektiviteten (KGE) for hvert nedslagsfelt og datasett. En progressiv Monte Carlo-metode innenfor en trinnvis kalibrering ble brukt. De resulterende utslippstidsseriene og vannbalansene i nedbørfeltene ble analysert og sammenlignet. I tillegg ble utvalgte nedbørshendelser, der nedbørsproduktene ikke var i stand til å beskrive atmosfæriske prosesser på riktig måte, analysert i dybden. Datasettene ble videre sammenlignet ved romlig akkumulering av årlige nedbørsummer over nedbørfeltene, ved å bruke en værstasjon eid av et vannkraftselskap for å evaluere passformen til dataene og ved å sammenligne avrenning og nedbørsmengde i bassengene.

Det viste seg at bruk av rå radardata ikke er rimelig på grunn av stråleskjermings- og dempningseffekter, spesielt i innlandet, samt en høy underestimering av nedbørsmengden. En modell med seNorge2 viste farbare resultater under kalibrering, men falt betydelig under validering. NWP så ut til å produsere mer stabile modellytelser og hadde bedre samsvar med hensyn til vannbalanse og deteksjon av småskala nedbør. Modellytelsen er imidlertid lavere i de fleste nedslagsfelt.

Under en sensitivitetsanalyse ble ulike modellusikkerheter funnet og diskutert. De inkluderte ufølsomme parametere som ikke konvergerte under kalibreringen. I tillegg viste veksling av kalibrerings og valideringstidsperiode ustabilitet i modellresultatene avhengig av den valgte tidsperioden i noen nedbørfelt. Individuelle parametere ble også evaluert. Noen av dem viste fysisk rimelige verdier, men for eksempel virket elvehastigheten til å være unaturlig høy i noen tilfeller.

Acknowledgments

This thesis was financed as part of HydroCen 3.5 (https://www.ntnu.no/hydrocen/5.-vannkraftressurser-og-flerbruksinteresser). This project aims to adapt existing tools to assess hydropower potential and to use them for environmental design.

Thanks to SINTEF for including me into the group and showing me the tasks and work processes in a research environment. I made a lot of new experiences and learned from joining meetings and project discussions. I admire the passion and knowledge that you bring into each of your projects and wish you all well for the future.

I want to send special thanks to my supervisor Lennart for supporting me, answering all my questions and for a lot of interesting discussions. Additionally, thanks for improving my work with all your suggestions. I wish you a great time as a PhD student and hope I will meet you again when you visit Stuttgart. I also would like to thank my examiner Jochen for the regular meetings, the great e-mail support and for helping me getting to Norway in the first place.

Thank you to my family for support me in every possible way and special thanks to my sister for correcting this thesis. I also want to thank my friend Carolina for her company and support during dark and cold Norwegian lockdown days and the nice time we had together. I hope we will meet again soon. Also thank you to Sahra and Asja for the great time and fun experience we had together in Trondheim.

Contents

List of Figures

List of Tables

Acronyms

AET	Actual Evapotranspiration
AROME	Application of Research to Operations at Mesoscale
CDF	Cumulative Distribution Function
CLC	CORINE Land Cover
CORINE	Coordination of information on the environment
DEM	Digital Elevation Model
DEMC	Differential Evolution Markov Chain
ECMWF	European Centre for Medium-range Weather Forecasts
EOSDIS	Earth Observing System Data and Information System
EPS	Ensemble Prediction System
GLUE	The Generalized Likelihood Uncertainty Estimation
GMAO	Global Modelling and Assimilation Office
HBV	Hydrologiska Byråns Vattenbalansavdelning
HYPE	Hydrological Predictions for the Environment
IDW	Inverse-Distance-Weighting
KGE	Kling-Gupta Efficiency
KliNoGrid	Norwegian Climate Gridded dataset
MAE	Mean Absolute Error
m a.s.l.	meters above sea level
MEPS	MetCoOp Ensemble Prediction System
MET	Norwegian Meteorological Institute
MetCoOp	Meteorological Cooperation on Operational Numerical Weather Prediction
MODIS	Moderate Resolution Image Spectroradiometer
MSE	Mean Squared Error
NSE	Nash-Sutcliffe-Efficiency
NWP	Numerical Weather Prediction
NVE	Norwegian Water Resources and Energy Directorate
PDF	Probability Distribution Function
PET	Potential Evapotranspiration
RADAR	Radio Detection and Ranging
RE	Relative Bias
RMSE	Root Mean Square Error
SLCs	Soil type Land use Combinations.
SMHI	Swedish Meteorological and Hydrological Institute
WS	Weather Station

1 Introduction

Creating accurate and applicable discharge information of river systems for forecasting, planning or hydro power assessment is the aim of hydrological models. In Norway, hydro power produces about 90 % of Norwegian total power production (Norwegian Ministry of Petroleum and Energy 2021). That's why it is important to provide accurate forecasts and assumptions of water availability. The main cause of runoff is precipitation. Since precipitation has a high spatial and temporal variability over a large range of scale (Berne and Krajewski 2013), it is necessary to provide measurement data that captures these variabilities. For distributed models and simulation of streamflow of hydrological models, accurate precipitation data is required (Syed et al. 2003). The uncertainty of simulated discharge can be caused by inaccurate precipitation data to a large extend (Bárdossy et al. 2022).

Precipitation is measured the most accurate by rain gauges (WMO 2008). However, they aren't free of measurement errors. Førland et al. (1996) stated that in Nordic countries precipitation is underestimated by ground measurements because of wind-induced errors. In wind-exposed coastal and mountainous areas, precipitation may be underestimated by more than 50 % in winter, and an underestimation will probably also occur in summer for unshielded gauges (Førland et al. 1996). Besides aerodynamical effects, evaporation from gauges, snow drift, splashing and leakage can also influence the accuracy of gauge data. Besides measurement errors, the density of gauges in Norway is scarce in mountainous areas and in the north (Lussana et al. 2018). That's why other precipitation data sources have to be considered when setting up a hydrological model.

In recent developments, radar measurements are increasing and can be used for hydrological models (Fitsum et al. 2013, Gonchukov et al. 2018). However, these measurements can have significant errors and limit hydrological application (Berne and Krajewski 2013). For example, when using Radio Detection and Ranging (Radar) various measurement errors like attenuation, clutters or beam shielding can appear. Still, radar data provides spatial precipitation distributions over a large range and can improve the precipitation estimation over a catchment, especially in data scarce regions. Another source of spatial precipitation distributions are numerical weather models describing fluxes in the atmosphere. These models can be reanalysed with observations and then be included into hydrological models.

This study aims to test and evaluate different input products of hydrological models. As a first step, the datasets were generally analysed by comparing annual precipitation sums, correlations, mass balances and the relation of the data with private weather stations. Since the ground truth is not known, hydrological observations can provide information about meteorological processes in the catchment and therefore verify observations and numerical models.

In this thesis, five catchments that are located in different hydrological regions in Norway were chosen for hydrological modelling. A Hydrological Predictions for the Environment (short: HYPE) model developed by the Swedish Meteorological and Hydrological Institute (short: SMHI) was employed. The model was calibrated with different precipitation datasets. An hourly gauge interpolated dataset (seNorge2), an hourly dataset based on a numerical weather model combined with gauge and radar data (MEPS-Nordic) and in some of the catchments gauge data were used to model the study sites. To create reproducible results, different calibration methods were tested focusing on the optimisation method, number of simulations, multi-gauge calibration and handling low flows. After calibration, the results were analysed by evaluating the model performance, water balance and looking at specific events. An inverse hydrological model, that was developed by Kirchner (2009), was applied for these events to compare simulated precipitation based on discharge data with the estimated precipitation from the datasets. At last, a sensitivity analysis was done for the model using a Generalized Likelihood Uncertainty Estimation (GLUE) method to test the model uncertainty and its equifinality (Beven and Binley 1992).

2 Theory

This chapter provides an overview over basics of hydrological models, calibration methods and objective functions as well as inverse hydrological modelling. Different measurement methods of precipitation are discussed as well as their spatial information. At last, studies where hydrological models were applied for testing precipitation datasets are described.

2.1 Hydrological models

Hydrological models are simulating the water cycle in a system using mathematical functions to reconstruct physical processes. In general, the runoff (*Q*) of a catchment is calculated as the difference between precipitation (*P*) and evapotranspiration (*E*) plus storage changes (*dS*).

$$Q = P - (E + dS) \qquad Eq.\ 1$$

Runoff and evapotranspiration drain the catchment, precipitation is the inflow. Storage changes are typically caused by changes in the soil moisture, snow accumulation, glaciers, or groundwater storages. Over long time periods the storage change is usually zero (Ojha et al. 2008).

Precipitation data is one of the most important inputs for a hydrological model (Wood et al. 2000). Usually, it is measured by point stations located in the catchment. Since the distribution of precipitation can have an influence on the hydrologic response (Yang et al. 2013) or is locally concentrated (Berg et al. 2016), there is a necessity of having accurate precipitation data, that also describes the spatial distribution. This can be archived by using remote sensing data (weather radars, satellites) or simulate the precipitation using numerical weather models.

Evapotranspiration leads to water loss in the system, because of water vaporizing from the soil, plants, lakes etc. There is a difference between potential evapotranspiration (PET), that only depends on climatic factors and actual evapotranspiration (AET), that is limited by the amount of available water in the system (Thornthwaite and Mather 1951). A possible measurement method for evapotranspiration is a Lysimeter, but usually the evaporation is calculated by using an empirical mathematical equation. PET is calculated by using climatic variables, for example temperature, radiation and wind speed (Hargreaves and Samani 1985, Allan et al. 1998). AET can be estimated within a model to take the water availability into account.

Discharge is a typical output of hydrological models. Measurement methods can be tracers, flow meters or ultrasonic sound, or the discharge can be calculated using rating curves (NVE 2015a). These curves can be applied in hydraulically controlled sections, where the relation between water level and discharge is known. Accurate discharge data is important for the calibration and validation of the model.

Typical applications of hydrological models are flood and drought forecasting, assessment of water resources, simulating scenarios of human impacts and hydropower modelling (Beck et al. 2017). A lot of different hydrological models have been developed over the years. In general, they are distinguished in deterministic and stochastic models. Deterministic models can be classified into empirical, conceptual, and physically based models: empirical models describe a precipitation-runoff relationship that is based on observations (Sitterson et al. 2018), conceptual models use observed empirical relationships to build a precipitation-runoff-model and physically based models use only physical laws to calculate the runoff (Liu et al. 2017). A deterministic model always calculates the same output, unlike a stochastic model, that has a partial randomness (Milad et al. 2012). Models can also vary in temporal (from event to decade) and spatial (small to global) scale. There are distributed, semi-distributed and lumped models. Lumped models average the catchment without taking spatial distribution into account, semi-distributed models divide the

catchment into a certain amount of sub-catchments and fully distributed models are divided into grid-cells (Sitterson et al. 2018).

2.1.1 Calibration and validation

A hydrological model consists of state or prognostic variables that define the processes within the catchment. To create a consistent model that is close to observed discharge measurements these parameters have to be calibrated. The resulting parameters are then applied on another time period, location or season to evaluate the applicability. This process is called validation. During calibration, an optimal parameter set is detected that results in the best agreement of simulated and observed discharge according to the objective function. Over the years, different criteria were developed that describe the performance of a model. They are important for the calibration as an index of agreement, since they decide which parameter set is assumed to be the best fit. The resulting parameter set can also vary depending on the objective function. It describes the goodness of fit of a model. There are different functions available, some of them are described here.

First introduced by Nash and Sutcliffe (1970), the Nash-Sutcliffe-Efficiency (NSE) is a commonly used criteria for quantifying model performances. NSE is the quotient of residual variance and initial variance. It is calculated by using simulated values (x_i), observed discharge values (q_i) and the average observed discharge ($\bar{q}$). The NSE tends to be influenced by outliers, time-offset bias and magnitude. High discharges have more influence on the NSE and low discharges get neglected, because of the squared difference (Krause et al. 2005). Still, it is a reliable goodness-of-fit (McCuen et al. 2006).

$$NSE = 1 - \frac{\sum_{i=1}^{n}(x_i - q_i)^2}{\sum_{i=1}^{n}(q_i - \bar{q})^2} \qquad \textit{Eq. 2}$$

Another criteria is the Kling-Gupta Efficiency (KGE) described in Gupta et al. (2009). KGE is a multi-objective criteria, where correlation (r), variability error (using the standard deviation σ) and bias error (using the mean μ) are considered. They are combined by applying Euclidian distance. The KGE can have values between -∞ and one, with one being a perfect model fit. The relation between NSE and KGE is non-unique and can therefore not be compared directly (Knoben et al. 2019).

$$KGE = 1 - \sqrt{(r-1)^2 + \left(\frac{\sigma_{sim}}{\sigma_{obs}} - 1\right)^2 + \left(\frac{\mu_{sim}}{\mu_{obs}} - 1\right)^2} \qquad \textit{Eq. 3}$$

Another performance measure is using the correlation between two datasets. Correlation quantifies the interdependence between variables. An example of that is the Pearson correlation (r). The Pearson correlation is a quantitative measure to determine the linear relationship between two variables (Boslaugh and Watters 2008) and has a value between -1 (negative correlation) and 1 (positive correlation). A value around zero represents a weak or non-linear relationship. It is calculated according to Eq. 4 with the sample size n, sample points x_i and the mean of the sample $\bar{x}$.

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}} \qquad \textit{Eq. 4}$$

The Pearson correlation gives no information about the proportion of variation (Boslaugh and Watters 2008). The coefficient of determination (r^2) however can provide that information by

describing how differences of one variable can be explained by another variable. For that, the Pearson correlation is squared.

$$r^2 = \left(\frac{\sum_{i=1}^{n}(x_i - \bar{x})(q_i - \bar{q})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(q_i - \bar{q})^2}} \right)^2 \quad \textit{Eq. 5}$$

Another possibility to calculate the correlation of a sample is by using the ranks of the variables. This approach is called Spearman rank order coefficient (r_s). This coefficient is less sensible towards outliers, because ranks instead of values are applied (Al-jabery et al.). It is calculated by ranking two samples (here: sample 1 and 2) and calculating the correlation of the ranks.
All these criteria are not scale dependent. That is important when comparing different datasets with each other that have different scales. The usage of scale dependant methods for calibration is possible as well. The most common used scale dependent measure is called Mean Square Error (MSE) or Root Mean Square Error (RMSE). RMSE is often preferred because it is on the same scale as the data (Hyndman and Koehler 2006).

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(x_i - q_i)^2 \quad \textit{Eq. 6}$$

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - q_i)^2} \quad \textit{Eq. 7}$$

Although they are widely applied, they are more sensitive towards outlier than the Mean Absolute Error (MAE) or the bias (Hyndman and Koehler 2006). The bias can be made scale-independent by calculating the relative bias (RE).

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|x_i - q_i| \quad \textit{Eq. 8}$$

$$bias = \frac{1}{n}\sum_{i=1}^{n}(x_i - q_i) \quad \textit{Eq. 9}$$

$$RE = \frac{\sum_{i=1}^{n}(x_i - q_i)}{|\sum_{i=1}^{n} q_i|} \quad \textit{Eq. 10}$$

2.1.2 Sensitivity and uncertainty analysis

Finding an individual set of parameters with the best model performance during a model calibration is no indication for the certainty of a model prediction. This is for example caused by the existence of many combinations of parameter sets that can result in the same model performance (Sorooshian 2008). An analysis to assume the uncertainty of parameters should be applied. In a sensitivity analysis, the uncertainty of a model input can be appointed to sources of uncertainty in the model inputs.
A sensitivity analysis is the critical examination of the relations between input and output of the model. The aim is to identify parameters that have a strong impact on the model results, as well as interactions between them and their spatial variability. There are different sources of uncertainty, like the model structure, parameter values, input data and boundary conditions (Beven and Binley 2014). A sensitivity analysis can help to identify potential deficiencies in model structure and

formulation, be a guide for parametrization and analyse the information given by observations (Devak and Dhanya 2017).

According to Devak and Dhanya (2017), different types of information can be extracted during a sensitivity analysis. One result is the determination of response surface, meaning to find the most influenced area depending on the model or the input data. A determination of the most influential uncertain parameters to rank the influence of the input parameters. Determining interaction between the parameters to find out the nature of processes. Finding non-influential parameters and remove them to reduce the model complicity.

There are different approaches for a sensitivity analysis. A simple approach is varying parameters one at a time. That means one parameter is changed with a specific criterium while all others are fixed. The model is run and the change of the output analysed. This method however does not provide information about combinations of parameters, especially in highly parameterized and complex models (Devak and Dhanya 2017).

For more complex systems with a higher interaction between parameters, the variance-based method can be applied. Here, the variance of the output is compared with the variance of the input while one parameter is fixed and the others are varied. The variance of the output data can then be an index of the sensitivity of that parameter (Saltelli et al. 2010).

Another approach was introduced by Beven and Binley (1992) and has been applied numerous times. The method is based on the observation that different parameter sets lead to equal results according to performance measures. This is called an equifinality of model structures or parameter sets that create equal results giving a non-uniqueness to parameter sets and models (Beven and Binley 2014). The Generalized Likelihood Uncertainty Estimation (GLUE) method combines calibration and uncertainty assessment and allows equifinality. The model outputs are weighted by their calibration period performance. The set of best predictions (behavioural model space) is then used for the expression of uncertainty. For that, the model is executed with different parameter sets and the variance of the results is used for uncertainty assessment (Sivasubramaniam et al. 2020). The parameter space of the behavioural model space can further be applied for a sensitivity analysis and to test equifinality. Wide varying parameters are assumed to be less sensitive than less varying parameters.

Another easy approach to test the reasonability of modelling results is to change calibration and validation time period to test the difference in the model performance. A model with accurate model structure and input data should not change the output drastically in that case.

2.1.3 Inverse hydrological modelling

In Kirchner (2009) a first-order nonlinear differential equation based on the water balance of a catchment (Eq. 1) is used for characterizing the basin. The precipitation (*P*), evapotranspiration (*E*) and discharge (*Q*) can provide information about the storage change (*dS/dt*). This means that the change of discharge is a function of storage. Recession curves can be used for quantifying the storage term.

$$Q = f(S) \qquad \textit{Eq. 11}$$

$$S = f^{-1}(Q) \qquad \textit{Eq. 12}$$

$$\frac{dQ}{dt} = \frac{dQ}{dS}\frac{dS}{dt} = \frac{dQ}{dS}(P - E - Q) \qquad \textit{Eq. 13}$$

Because the storage is a function of Q (see Eq. 12), the derivative dQ/dS can be described as a function of Q. This function is called the sensitivity function of the catchment g(Q). By combining Eq. 1 with the derivative dQ/dS, the following relation can be created.

$$g(Q) = \frac{dQ}{dS} = \frac{\frac{dQ}{dt}}{\frac{dS}{dt}} = \frac{\frac{dQ}{dt}}{(P - E - Q)} \qquad \textit{Eq. 14}$$

Eq. 14 can be simplified when evapotranspiration and precipitation fluxes are low compared to discharges. These events occur for example for no precipitation and in the night time.

$$g(Q) = \frac{dQ}{dS} \approx \left.\frac{-\frac{dQ}{dt}}{Q}\right|_{P \ll Q, E \ll Q} \qquad \textit{Eq. 15}$$

This function depends only on discharge measurements. The sensitivity function can be estimated by plotting the flow recession rate (-dQ/dt) as a function of Q (see Figure 1). The differential equation dQ/dt can be solved by assuming the flow recession rate to be a power law ($-dQ/dt = aQ^b$). With empirical measurements, g(Q) can be fitted according to the function:

$$\ln\big(g(Q)\big) = ln\left(\left.\frac{-\frac{dQ}{dt}}{Q}\right|_{P \ll Q, E \ll Q}\right) = c_1 + c_2 \ln(Q) + c_3 (\ln(Q))^2 \qquad \textit{Eq. 16}$$

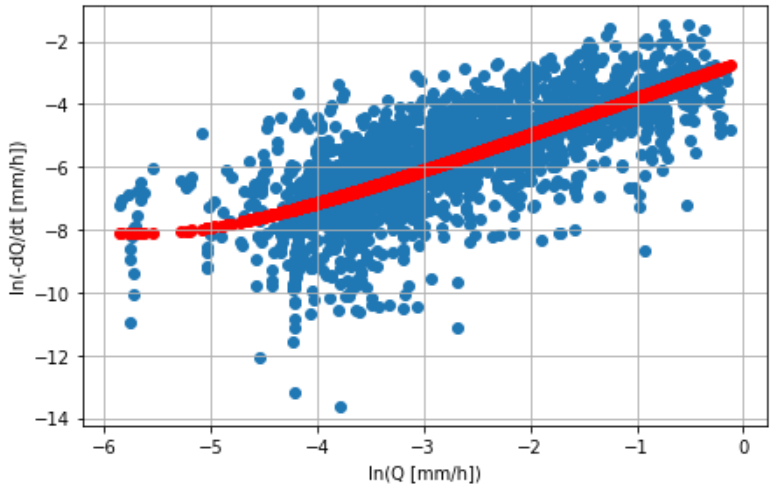


Figure 1: Example of a sensitivity curve with c_1=-2.6, c_2= 1.2, c_3= 0.004

Combining Eq. 13 and Eq. 14 yields to Eq. 17. This relation can be applied by calculating the discharge change of the next timestep with the discharge value of the previous one.

$$\frac{dQ}{dt} = \frac{dQ}{dS}\frac{dS}{dt} = g(Q)\left(\frac{P - E}{Q} - 1\right) \qquad \textit{Eq. 17}$$

The advantage of this method is that discharge data is directly used for the hydrological model, instead of being only included for calibration or certainty analysis. Discharge does not depend on spatial variability (in contrast to precipitation), because it is an accumulated measurement from the upstream area. Another advantage of this approach is that it can be inverted. This means a

"backward hydrology" can be created, where precipitation is calculated only based on discharge data and without calibration. For that, Eq. 17 can be rearranged to:

$$P - E = \frac{\frac{dQ}{dt}}{g(Q)} + Q \qquad \textit{Eq. 18}$$

In a precipitation event, the air humidity is saturated, resulting in a small evapotranspiration. That's why it is assumed to be zero. In a discretised manner, the equation looks like this:

$$P_t \approx max\left(0, \frac{(Q_{t+1} - Q_{t-1})/2}{(g(Q_{t+1}) + g(Q_{t-1})/2} + (Q_{t+1} - Q_{t-1})/2\right) \qquad \textit{Eq. 19}$$

2.2 Point and spatial precipitation data

As mentioned in the introduction, input data of hydrological model is the main source of uncertainty. This uncertainty is caused by measurement errors, unknown spatial variability and data scarcity. Especially precipitation data is a source of error, because of its high spatial and temporal variability as well as measurement error sources. In this section, different methods to measure precipitation and to assume its distribution are described.

2.2.1 Gauges and point station interpolation

A precipitation gauge is an instrument that directly measures the precipitation amount of unit area per time. It consists of a collection container that is placed in an open area (Acharya 2017). Precipitation is measured as height of a volume. That's why the area of the container is not important, but should not be too small (Acharya 2017). The data can be received manually or automatically. In the past, the precipitation amount was manually obtained by checking the container and measuring the filling of it. Nowadays, this can be done automatically. Examples are tipping bucket rain gauges, where a bucket is filled until a threshold is achieved and the bucket tips creating an electrical signal (Acharya 2017). Other examples are weighting the precipitation amount inside the collector and calculating the precipitation height from weight changes during precipitation events or counting the number of drops by a photoelectric barrier. The measurement of snow is achieved by using collectors with heating function, so that the snow melts and is measured the same way as rain.

Although gauge data is assumed to be accurate (WMO 2008) there are still some measurement errors because they are used in the open field. Some of them are discussed in the following. Detention water sticking at the inner wall of the collector and evaporating there can lead to an underestimation of precipitation. Additionally construction errors or wrong calibration of the gauges lead to an bias in the measurements (Sevruk 2004). Other errors can occur from influences of the nearby environment, for example leaves from trees, insects or walls. The highest source of error is however wind-induced under catch of precipitation (Sevruk 2004). An under catch is caused by turbulent wind fields above the gauge leading to a drift of rain drops and snowflakes instead of going inside the collector. Wolff et al. (2015) measured significant wind under catch in a field study in Norway. Especially for snow and high windspeeds, the under catch can be up to 80% of the precipitation amount. Rain measurements are less affected by wind. A relationship between windspeed, temperature and under catch was developed by Wolff et al. (2015). Figure 2 shows an example of under catches influenced by wind speed and temperature.

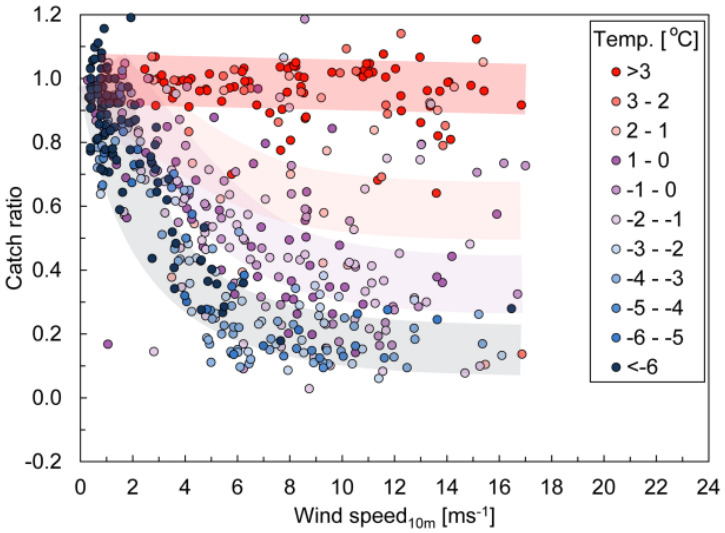


Figure 2: Under catch of precipitation dependent on the temperature and wind speed (Wolff et al. 2015)

Besides measurement errors, a point measurement does not provide information about the spatial distribution of the precipitation. Precipitation fields are of interest in hydrological models to calculate the areal precipitation amount. That is why different methods were developed that create spatial interpolated precipitation fields.

When the areal precipitation (h_N) is of interest, the easiest approach is averaging nearby gauges ($h_{N,i}$) with weighting them all equally (Eq. 20). Another approach, called Thiessen Polygon, defines an area inside the basin (A_i) for every station. These areas can then be used as a weight for the areal precipitation (Eq. 21). Another method is interpolating the precipitation by using the distance (d_i) of the station to the catchment or a grid cell in the Inverse-Distance-Weighting (IDW) (Eq. 22). This method is one of the most used deterministic interpolation methods because of the easy implementation (Ryu et al. 2021).

Averaging $$h_N = \frac{1}{n}\sum_{i=1}^{n} h_{N,i} \qquad \textit{Eq. 20}$$

Thiessen Polygon $$h_N = \sum_{i=1}^{n} \frac{A_i}{\sum_{i=1}^{n} A_i} h_{N,i} \qquad \textit{Eq. 21}$$

IDW $$h_N = \sum_{i=1}^{n} \frac{\left(\frac{1}{d_i}\right)^2}{\sum_{i=1}^{n}\left(\frac{1}{d_i}\right)^2} h_{N,i} \qquad \textit{Eq. 22}$$

Another possibility for interpolating precipitation data is using geostatistical approaches. These methods can take catchments characteristics into account by analysing the spatial variability in so-called variograms. The main idea is that the variance between two nearby points should be smaller than points that have a higher distance (Sen 2016). It can be calculated by summing up the values (Z_i) squared difference between points with a specific distance (d). The number of values with that distance is n_d. The semi-variogram equation is shown in Eq. 23.

$$\gamma(d) = \frac{1}{2n_d}\sum_{k=1}^{n_d}(Z_i - Z_{i+d})^2 \qquad \textit{Eq. 23}$$

Kriging is a geostatistical interpolation technique that depends on distance and variation between known points to calculate estimates of unknown points (Ryu et al. 2021). An empirical semi-variogram function (γ) has to be fitted for that process. In the kriging process, weights for each surrounding point are calculated by using the fitted semi-variogram. It produces the best linear unbiased estimator of the values at locations with no information of the value (Zawadzki et al. 2005).
Another method is optimal interpolation. Here, observation data is paired with a background field, for example the topography or climatological fields. It is very similar to kriging (Sen 2016). The weighting coefficients get determined by minimizing the least square fit between point information and background field.
A background field is also acquired in the Bayes Interpolation. In this method, the likelihood of the interpolated field (I) to the observation (O) and the background field (B) is maximized. The posterior probability $\mathrm{P(O|I)}$ must be maximized by maximizing the likelihood $\mathrm{P(I|O)}$ and prior $\mathrm{P(I)}$. To increase the likelihood, the interpolation has to be close to point observations, to increase the prior, the interpolation has to be close to the background field.

$$P(O|I) \sim P(I|O)P(I) \qquad \textit{Eq. 24}$$

2.2.2 Weather Radars

Precipitation is heterogenous in time and space. Gauge stations measure point information about precipitation on the ground with sufficient accuracy for modelling purposes. Yet, when looking at a wide-scale catchment, the precipitation amount is not homogeneous. To take this into account, a dense gauging network is necessary. Since this cannot be provided everywhere (because of e.g. costs, data collection and infra structure), other spatial information has to be provided. Weather radars provide an overview over a wide scale in a high spatial and temporal resolution. They are an important tool for assessing the precipitation distribution (Collier and Hardaker 2004) and are capable to detect severe weather and precipitation over a wide scale.
RADAR is an acronym for Radio Detection and Ranging and is an indirect measurement method, where radio waves are emitted by a transmitter and reflected by an object in the atmosphere. The reflected energy “echo” is then measured by the receiver (Raghavan 2003). According to the direction and time of travel of the echo the position of the object can be detected. If a target is moving radial to the radar, the reflected radiation will have a shifted frequency depending on the velocity of the target (doppler shift) (Gekat et al.). The radiation can be sent with horizontal or vertical polarization. The echo will then have the same polarization. When using both polarizations, an information of the spherical shape of an object can be archived (Raghavan 2003). Polarized radars can detect different shapes, orientations and dielectric constants of the rain drops and classify different rain types to improve the quantitative rain estimate (Gekat et al.).
Besides the location of precipitation, the radar reflectivity (*Z*) provides quantitative information about the amount of water. Z is in relation with the drop diameter and the number of drops. Since the drop size distribution is usually not known, empirical laws were developed to calculate the precipitation rate (*R*). One possible relation was developed by Marshall and Palmer (1948) and uses a power law to describe the relation between reflectivity and rain rate (R).

$$Z = 200 \cdot R^{1.6} \qquad \textit{Eq. 25}$$

This empirical relationship depends on the drop size distribution, as well as the phase of the precipitation and should be used cautiously (Raghavan 2003, Fujiyoshi et al. 1990, Joss et al. 1990). The interest in spatial distributed precipitation data is increasing, for example for accurate weather forecasts or to be used in hydrological models. Hydrological models nowadays are more spatially distributed and need spatial information of the precipitation (Berne and Krajewski 2013). Besides that, flash floods usually occur in small spatial scales and are often not measured adequately from gauge networks (Poméon et al. 2020 and Šálek et al. 2006). Radar data is more likely to capture such events. Lengfeld et al. (2020) found out, that only 17.3 % of hourly heavy precipitation events are captured by gauges in Germany compared to radar data. Another important application is urban hydrology where a high resolution of precipitation data is needed (Berne et al. 2004 and Einfalt et al. 2004). Because information about spatial distribution in the catchment area is usually not available in a comparable way from gauge stations (Syed et al. 2003), radar data can be used to fill information gaps. A possibility is improving the quantitative estimation by merging gauge data with radar measurement for spatial interpolation (Hasan et al. 2016, Sivasubramaniam et al. 2019). Furthermore, radars can provide precipitation measurements in ungauged areas (Berne and Krajewski 2013).

Besides the advantages of weather radar data, the main disadvantage is erroneous data. Since the data is detected indirectly, many errors occur that need to be corrected. The most challenging ones are described in this paragraph. **Clutters** occur when stationary objects, for example buildings, mountains or the ground reflect the radiation and create an echo that is detected as precipitation (Gekat et al.). They can usually be corrected with “clear-sky” images. When travelling through the atmosphere, the radiation suffers attenuation. **Attenuation** depends on the precipitation amount as well as the precipitation state and is complicated to correct systematically (Gekat et al.). Another error can occur at the melting layer of a precipitation cloud. This layer is called the **bright band**. When snow is melting in that layer, a water film is covering the snowflake. The mix of water outside and a large diameter of the snow flake cause a high reflectivity that is detected by the radar. The **height of the beam** is increasing with range. This can cause missed precipitation events or measuring precipitation that evaporates before arriving at the ground. To avoid ground clutter, the radar beam is usually in an elevation angle. This causes a high altitude of the radar beam at higher ranges and leads to differences between the measured height and the ground level (Berne and Krajewski 2013). In mountainous regions, effects like **shielding**, **partial beam filling** and **beam-broadening** are dominant sources of error (Gekat et al.). Here, the radar beam is blocked in some range gates and the precipitation behind is not measured accurately (Berne and Krajewski 2013). **Anaprops** can occur when there are anomalous atmospheric conditions that change the propagation of the radar beam and result in wrongly located radar echos (Berne and Krajewski 2013).

Because precipitation is not measured directly from radars, another source of uncertainty is the relation between radar reflectivity and rainfall rate (Berne and Krajewski 2013). As mentioned before, the **Z-R relation** is empirical and does depend on the phase of precipitation, as well as the drop size distribution. Figure 3 shows examples of measurement errors of radars. The figure additionally shows errors like a wet radome, ships, aircrafts, chaffs, insects, and wind turbines that wrongly get classified as precipitation.

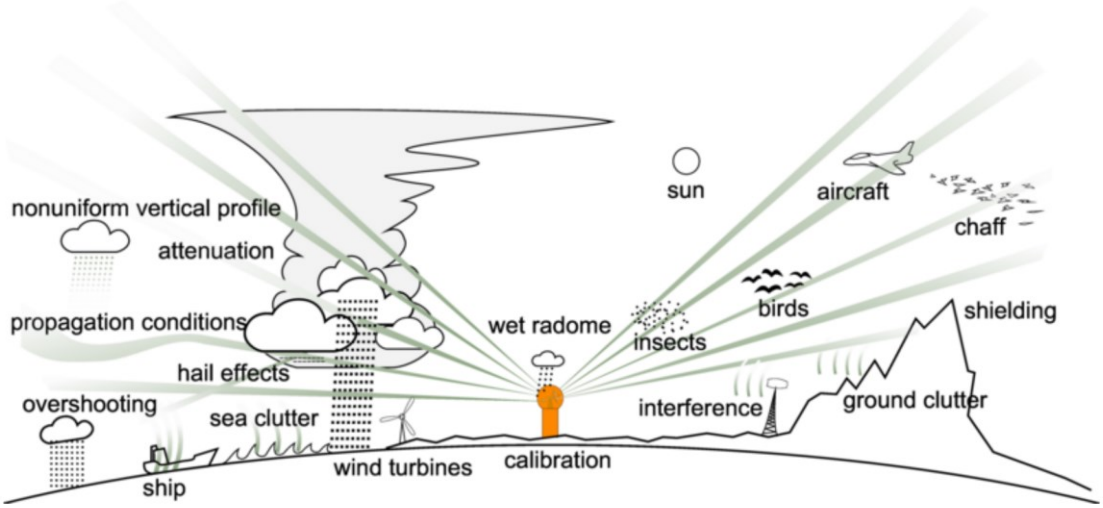


Figure 3: Examples of different radar measurement errors (Peura et al. 2006)

2.2.3 Numerical weather prediction (NWP)

Atmospheric processes are complex and non-linear. Numerical weather prediction (NWP) models estimate physical processes of the atmosphere and on the ground as well as their influence on air pressure, temperature, wind, clouds and precipitation. These physical processes can be described by using equations of fluid mechanics. The physical laws are based on energy, mass and momentum conservation (Kalnay 2012). The equations are differential and can only be solved in a numerical way by discretising them in space and time. A higher resolution increases the accuracy of a model (Kalnay 2012) but also increases the computing effort. To assure the stability of a model, the spatial and temporal resolution must be chosen in a reasonable matter.

NWPs are divided into deterministic and stochastic models. Deterministic models calculate a single forecast with a given set of initial state variables and parameters. Stochastic models create a number of simulations with varying initial conditions (Auer 2018).

An important task is to use the right initial conditions from observational data, because they are very sensitive in the model (Kimura 2002). Uncertainties in the initial state can lead to a high error in the forecast after some days (Wernli 2011). This is a major uncertainty, because there is never a full coverage of atmospheric variables available (Wergen 2002). Even for remote sensing data, that covers the area, the measurement can be biased or erroneous. Exact long-term predictions are not possible, because atmospheric flow is chaotic and not yet fully understood (Lynch 2008). Another challenge is to model convective precipitation, because of the small scale and the difficult forecast of these events (Lussana et al. 2021).

Uncertainties can be considered by using ensemble prediction systems (EPS). A set of forecasts instead of a deterministic one is calculated here. With that, different scenarios of forecasts are created. If all scenarios behave in a similar way, the uncertainty of the forecast is low (Wernli 2011) and the other way around. The aim is to quantify the uncertainty of the forecast because of erroneous initial conditions or modelling errors. The different scenarios can be used for forecasting as well as quantifying the probability of a forecast (Wernli 2011).

An example of an EPS is shown in Figure 4. Here, the initial state is assumed to be a probability distribution function (PDF). The numerical model then calculates different scenarios using different initial states according to the PDF. These ensembles are the thin black lines in the figure. Deterministic forecasts (thick black line) and actual observations (dotted line) are not similar in this example, whereas three of the ensembles show a more similar behaviour to the truth.

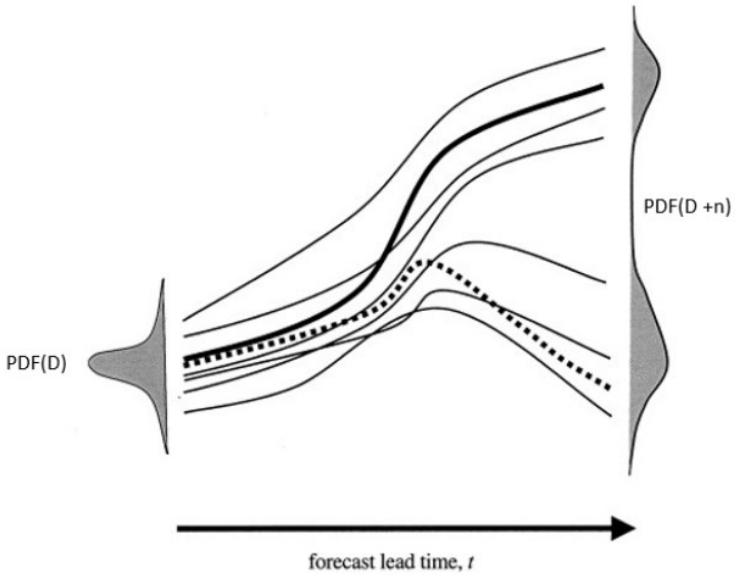


Figure 4: Example of EPS forecast (Lynch 2014)

All models are based on the same equations, but can vary in mathematical formulation and discretization scheme (Kalnay 2012). An example is the model developed by the European Centre for medium-range weather forecasts (ECMWF), that solves differential equations with a semi-langrangian, semi-implicit time scheme. The model has a grid size of 9 km and describes weather processes all over the globe. An ocean wave model is also included. The initial state is given by conventional and satellite data (Lynch 2008). Another model is the Application of Research to Operations at Mesoscale (AROME) developed by Météo-France in 2008 (Seity et al. 2011). It has a resolution of 2.5 km and covers Europe. The equations are solved in a semi-langrangian, semi-implicit discretization scheme using finite differences. The resolution was chosen to be able to predict thunderstorms, that are usually in a range of three to six kilometres (Seity et al. 2011). Initial data is observed by for example radiosondes, wind profilers, aircraft reports, satellites and doppler radars (Seity et al. 2011). AROME was further applied on the Nordic countries by creating the AROME-MetCoOp (Meteorological Cooperation on Operational Numerical Weather Prediction operational weather) forecast (Müller et al. 2017). This model has a resolution of 2.5 km as well. To adapt it to the Nordic region, it was calibrated with radar reflectivity data, ground-based satellite observations and gauge observations (Müller et al. 2017).
Possible applications of NWPs are short-range forecasting, warning of weather extremes and floods, marine guidance and predications of road ice (Lynch 2008). When modelling results get paired with observations, they are called reanalysed data. This data usually uses the model as a background, but takes observations (radar, gauge, satellites) into account, creating a spatial meteorological product. According to Lynch (2008), the best results can be achieved by combining NWP, observations (station, radar, satellite) and other data. Frogner et al. (2019) added that for smaller scale precipitation, post-processing should be done before using the model data. Reanalysed products can then be applied e.g. for hydrological models and risk assessments (Sivasubramaniam et al. 2020).

2.3 Evaluation of precipitation data with hydrological models

Using hydrological models to evaluate precipitation datasets is an often applied approach. Examples will be discussed in this section. Besides the hydrological model, the choice of meteorological data has a considerable influence on the model performance, as te Linde et al. (2008) concluded in their study. If the ground truth is unknown, the application of hydrological models can be a good possibility to evaluate spatial distributed precipitation products. This can be done by looking at water balances (evapotranspiration, storage changes and discharge) or analysing the model performance.

By creating a lumped hydrological model over the USA for 424 catchments, Essou et al. (2016) compared four climatic datasets. All these datasets were based on gauge-derived data with different interpolation methods. They found significant differences for the datasets, even though all of them were based on gauge data. Nevertheless, the impact of the datasets could be equalized during the calibration. Vaze et al. (2011) compared three different rainfall inputs with four different rainfall runoff models in 240 catchments (50-200 km^2) in the southeast of Australia. The rainfall data was extracted from single rainfall stations as point information, interpolated Thiessen polygons, and as average of interpolated rainfall areas. The models were lumped. The results showed a better performance in the model when using input data with better spatial representation of rainfall. Ledesma and Futter (2017) compared instrumental and gridded spatial interpolated precipitation and temperature datasets by modelling six catchments (0.037 to 725 km^2) across Sweden using two different hydrological models. Each model was calibrated separately. According to the study, gridded dataset resulted in a higher Nash-Sutcliff-Efficiency.

All of these studies showed a better performance of hydrological models with spatial interpolated rain gauge datasets. Nowadays, more information about the spatial distribution is known because of remote sensing technics and numerical models. Especially in data scarce regions, remote sensed or reanalysed data is the main source of meteorological information. For that reason, the data should be evaluated in better measured areas before applying them to data scarce regions.

Radar measurements have increased in the past and can be used for hydrological models (Fitsum et al. 2013, Krajewski and Smith 2002). However, these measurements have significant errors and can limit hydrological application (Berne and Krajewski 2013). Usually, radar data leads to an underestimation of discharge when applied to a hydrological model. For example Gonchukov et al. (2018) used radar precipitation for modelling a flash flood in Russia. Uncorrected radar data lead to an sever underestimation of discharge. The model was improved by bias correcting the precipitation. The combination of radar and rain gauge data can improve the precipitation significantly (Boudevillain et al. 2016, Borup et al. 2016, Foehn et al. 2018). Yu et al. (2015) further combined radar data with NWP because of the mountainous area in Japan causing beam blockage of radar beams. They concluded that this combination creates reasonable results when modelling flood events.

In Norway, different gridded datasets are available, including spatial interpolated, radar and numerical models. In the following section, different approaches to compare these products are summarized. In a previous SINTEF project, Abdella et al. (2012) tested three different precipitation products (gauge data, radar data and a combined product) by using a precipitation-runoff simulation. They examined daily radar-derived data, interpolated station data and a combination of those two. The simulation was done in 50 catchments at two radar range areas of Rissa and Hægebostad. That precipitation was fed into a distributed and a lumped model. They detected that the combined products calculated the best results based on NSE, bias and correlation coefficient. Still, all the input products led to an underestimation of runoff. The highest underestimation was found when using radar data. These results showed a bias in gauge and radar data, for example caused by beam blockage, under catch and a wrong Z-R relation.

Another model based on radar data was used by Reinemo (2012), that employed a Mike Urban model of the city of Trondheim during a flood in August 2007. Since extreme weather events are usually very variable over space and time, radar data with a resolution of 250x250 meters was used in a temporal resolution of one minute. The results showed a better simulation of critical areas with radar data than when using gauge data. The reason for that was the high local variability of the rain event.

Sivasubramaniam et al. (2020) made a case study comparing hydrological models with different precipitation input datasets. The three study sites had an area between 200 and 700 km^2. They compared precipitation from gauge measurements, interpolated station data and precipitation calculated using a numerical weather model. A time interval of three years for calibration and two years for validation was chosen. When comparing the correlation between model-based and gauge precipitation, it was found that the hourly values are variable, whereas the daily sum showed better results. The model performance with model-based precipitation showed comparable results to gauge and gridded datasets. The simulated flow volume also showed comparableness.

A different approach was used by Hailegeorgis et al. (2016), were a “top-down” model invented by Kirchner (2009) was implemented in the Gaula catchment, Norway. This was done with an hourly timestep and a distributed model in a mountainous area with a catchment area of over 3000 km^2. They used precipitation, wind speed and humidity data from gauge stations. The study showed acceptable results of these catchment in calibration and validation. However, the precipitation was not calculated as an inverse approach in this study.

In all of these studies, meteorological data was compared by using hydrological models and different input datasets. Most of these studies calibrated the catchment for each dataset individually. In general, spatial information seems to improve the model results, even with interpolated gauge data. Radar data can be used for hydrological modelling. It is especially an advantage for small-scale precipitation events. Still, radar data tends to underestimate the precipitation amount. A combination of gauge and radar data leads to better modelling results. Numerical model data can also be used as input data, because comparable results were created with them.

3 Material

This chapter includes an overview over the hydrological model that was applied for comparing precipitation datasets. The model was executed for five study sites in Norway. Their characteristics and location are described in the second section. A description and summary of the meteorological and hydrological measurements and products available in Norway is given in the last section.

3.1 HYPE model

HYPE was developed by the Swedish Meteorological and Hydrological Institute (SMHI 2021). It is applied for small and wide scales, even a land surfaces worldwide covering version was implemented by the SMHI (Arheimer et al. 2020). The main idea is to divide landscapes into different classes according to soil type, land use and altitude (Lindström et al. 2010). It was developed due to a lack of spatial distributed models with simple implementation and ability to simulate water quality (Lindström et al. 2010). HYPE is open source with the aim to be as simple as possible to be more flexible towards changing the source code. In the following chapters, HYPE is briefly explained. More information can be found on the HYPE documentation website (SMHI 2021).

3.1.1 Model structure and input parameters

The model domain consists of different sub-basins which are divided into classes. These classes are not dependent on location. It is possible to divide one basin with different land uses and soil types into a number of classes (Lindström et al. 2010). These classes are called “soil type land use combinations“ (SLCs). They can be compared to hydrological response units in HBV (SMHI 2021). In the Norwegian HYPE model first implemented by Schönfelder (2017) seven SLCs were defined. These classes are water, mountain, forest, marsh, glacier and a combination of urban and agriculture land use.

Water is divided into two special classes called *ilake* and *olake*. An *ilake* is defined as a lake inside a sub-catchment, that can store and evaporate water according to PET. An *olake* is located at the end of a sub-catchment (or a whole sub-catchment is defined as an *olake*). That way a lake has a high influence on the outflow discharge because of for example damping a flood curve. Discharge of *olakes* is implemented by using a water level-discharge relation. Parameters of this relation can be calibrated for the whole catchment (with possibly more than one lake) or added for each lake itself as input data. Another distinction in HYPE is made between main rivers and local streams. Main rivers concentrate discharge coming from local streams in the sub-catchment. Main rivers also include inflow of water from the upper catchments.

It is possible to model up to three soil layers in one SLC with variable thickness in each layer. Water retention can be varied by the parameters *wilting point, field capacity* and *effective porosity* (SMHI 2021). Water retention is allocated evenly between soil layers dependent on their thickness, or these parameters can be defined per soil layer. In this thesis, soil and land use were assumed to be similar, meaning that a specific land use has the same soil layers in each region and not a combination of different soils.

The runoff in HYPE is divided into groundwater runoff, depending on effective porosity, water level, surface runoff, when the water level is above the upper soil layer, and runoff through drainage pipes, which was not included in this thesis. Another water flow is infiltration of rain and snow melt into the upper soil layer and percolation between soil layers, depending on their field capacity (SMHI 2021). A summarizing flow chart can be found in Appendix A, Figure 46.

3.1.2 Automatic calibration

HYPE provides nine different automatic calibration methods. They are based on either Monte Carlo, Differential Evolution Markov Chain, Brent, method of steepest decent or quasi-Newton (SMHI 2021). Two of them, that showed a good efficiency and model performance in Schönfelder and Baclet (2022), are described in this chapter.

3.1.2.1 Differential Evolution Markov Chain

The Differential Evolution Markov Chain (DEMC) is a combination of a Differential Evolution method (Storn and Price 1995), an algorithm searching for a numerical optimum between parameter spaces (Braak 2006) and the Markov Chain Monte Carlo method, where random samples are generated, that depend only on the previous sample and are getting accepted or rejected, depending on their Metropolis-Hastings ratio (van Ravenzwaaij et al. 2018). The Metropolis-Hastings ratio is calculated by using a fitting function (π) of new and previous samples. The DEMC method fixes the problem of deciding jump sizes of the Markov Chain by including differences between two random parameter vectors (Braak 2006). In HYPE, the following parameters for DEMC are implemented.

- DEMC_ngen: number of generations
- DEMC_npop: size of generations
- DEMC_gammascale: determines how much the randomly generated dataset contributes to the new parameter set
- DEMC_sigma: derivation of random parameters
- DEMC_crossover: probability to not use mutated parameters
- DEMC_accprob: decides whether only better results are saved

The number of simulations depends on the number of generations and size of these generations (N = DEMC_ngen x DEMC_npop). In Figure 5 the process of the DEMC calibration is visualized.

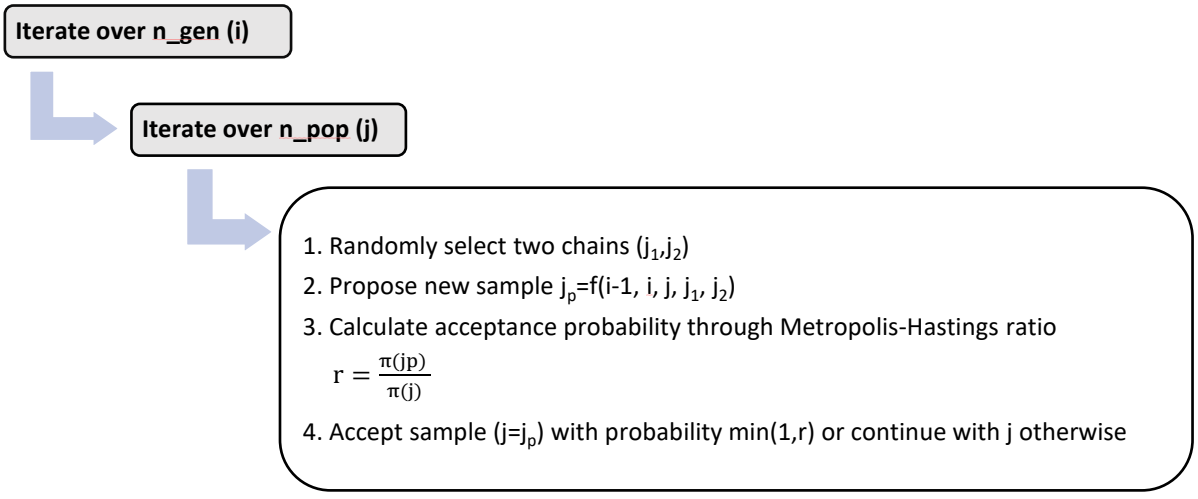


Figure 5: simplified DEMC process according to the algorithm provided in Braak (2006).

3.1.2.2 Progressive Monte Carlo Method

Progressive Monte Carlo is a directed Monte Carlo approach that takes the direction of the best results by reducing parameter spaces during calibration. For that, a certain amount (num_bpmc) of random parameter values are generated and simulated. After that, the best n numbers of results

(num_ens) are saved and the parameter space is reduced according to parameters of the best simulations. The next stage begins with reduced parameter space and again a certain number of random generated parameters are simulated. This is repeated until the maximum number of stages (num_bpmax) is reached (SMHI 2021).

- num_ens: number of best runs that are retained to form the next stage
- num_bpmc: model simulations per stage
- num_bpmax: number of stages

The number of calibrations is calculated as a product of stages times number of simulations in each stage (N = num_bpmc x num_bpmax). Figure 6 shows an example of this calibration process.

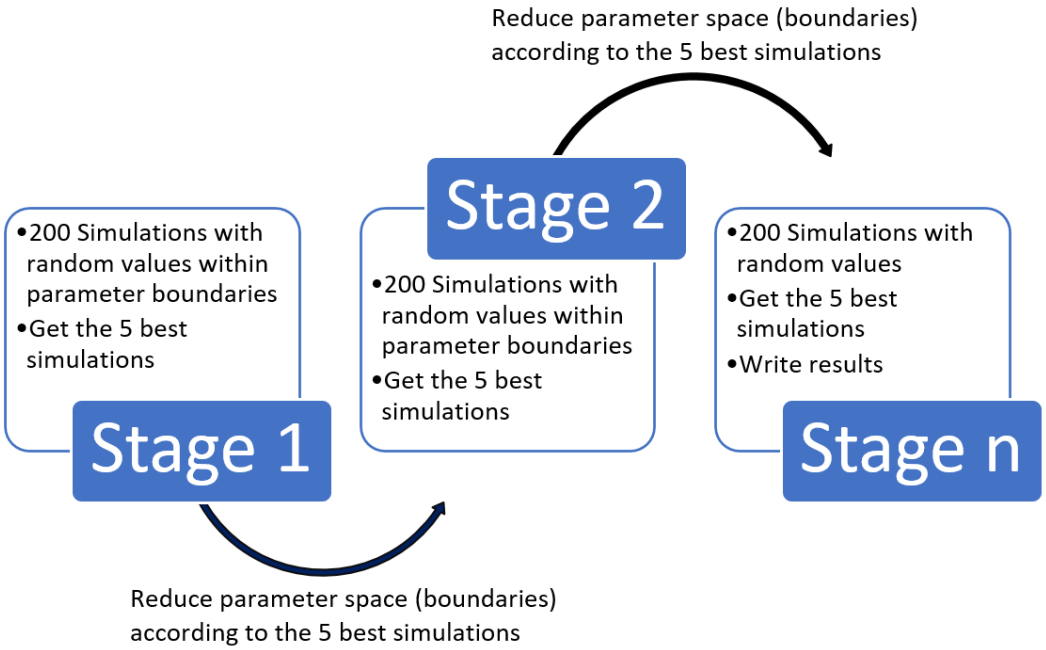


Figure 6: process of Progressive Monte Carlo, num_ens = 5, num_bpmc = 200, num_bpmax = n

3.2 Study areas

This study consists of five test cases where HYPE was applied. These catchments were selected with different criteria. Excluding catchments that are mostly influenced by hydropower, the remaining basins needed to have existing discharge measurement stations that have hourly data on the same time interval as the input data available. Another criterion is the amount of precipitation gauges near the catchment and the nearby radar stations. They also vary in size and altitude. A summary of characteristics and land uses of the five catchments is given in Table 1.

Table 1: Catchment characteristics and land uses

Catchment	Area	Mean Elevation	Min Elevation	Max Elevation	Mean slope	River length
	km^2	m.a.sl	m.a.sl	m.a.sl	°	km
Nausta	278	608	0	1467	16.7	561
Gaula	3659	674	0	1319	7.74	903
Usma	42	735	498	1058	10.6	111
Surna	1221	648	0	1664	12.0	263
Grunnåi	85	960	193	1411	14.4	27
SLCs	**Water**	**Mountain**	**Forest**	**Marsh**	**Glacier**	**Urban/Agriculture**
	%	%	%	%	%	%
Nausta	2.7	24.0	21.4	42.6	0.0	9.3
Gaula	2.	21.4	25.9	43.0	0.0	7.7
Usma	1.2	42.4	12.9	43.5	0.0	0.0
Surna	2.9	26.2	24.1	39.2	0.1	7.6
Grunnåi	3.1	24.2	31.8	40.5	0.0	0.4

To have a representative study over different regions in Norway the catchments were compared to a study of Gottschalk et al. (1979) that divided Norway into different hydrological regions. According to this study, Norway consists of six regimes. One of them being a mountain regime (H_1L_1), with high discharges in spring caused by snowmelt and low discharges in winter because of snow accumulation Another regime is called Atlantic regime (H_3L_3) which is more influenced by rainwater instead of snowmelt and has low flows during summer caused by less precipitation and a higher evapotranspiration. The inland regime (H_2L_1) is a transition zone between dominant rain and snowmelt discharges with low discharges in winter. There are three other regimes, two of them being transition regimes and one is the Baltic regime (H_2L_3) with a mix of rain and snowmelt water in spring and low flows in summer. The position of the regimes and locations of the catchments are visualised in Figure 7. The categorization of the test catchments in this thesis to the locations is summarized in Table 2. In the following sections, each catchment is described briefly.

Table 2: Location of the catchments according to their hydrological region (Gottschalk et al. 1979)

Regime	Catchments
mountain regime (H_1L_1),	Gaula, Usma, Surna, upper basin of Nausta
inland regime (H_2L_1)	Grunnåi
Atlantic regime (H_3L_3)	Lower basin of Nausta
Baltic regime (H_2L_3)	Parts of Gaula and Surna
transition regimes (H_2L_2)	-
transition regimes (H_3L_2)	-

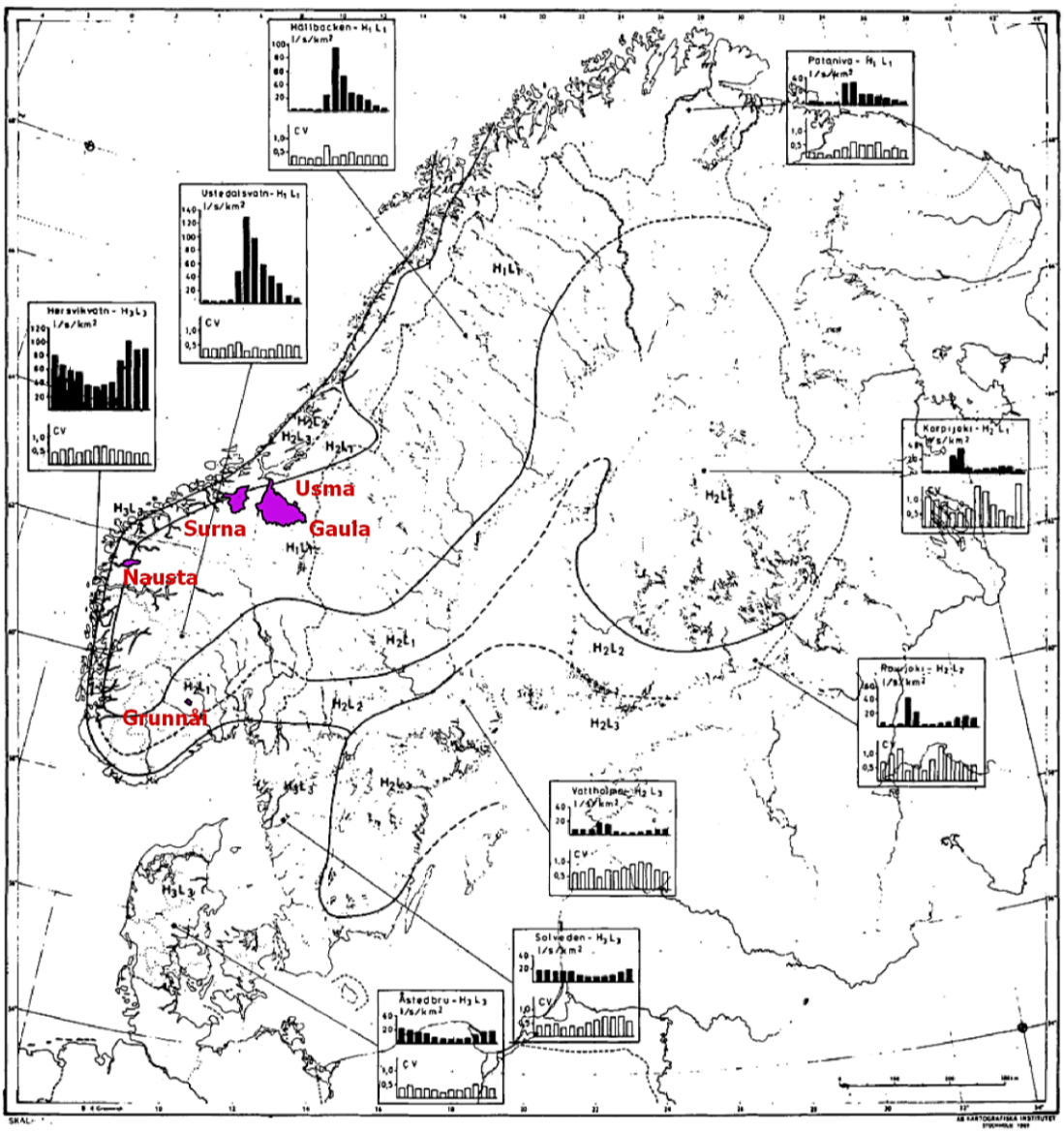


Figure 7: Hydrological regions in Norway according to Gottschalk (1979) and locations of the catchments. Map taken from Gottschalk (1979)

3.2.1 Nausta

Nausta is located at the upper west coast of Norway. With an area of 277 km^2 and an elevation between 0 and 1,467 m above sea level (a.s.l.), with a mean elevation of 608 m a.s.l., it is relatively steep. This catchment is unregulated, except for some small hydropower stations, that don't influence the hydrology in a significant way. For that reason and its location at the Atlantic coast, Nausta was added as a study area in this thesis.

A main part of Nausta is marsh area (43 %), with mountain areas at the edges of the basin and forest, agriculture land and urban areas located at lower elevated areas near the Nausta river. These areas are mostly crop fields with houses spread unconcentrated over the area. There is one city located at the river mouth. According to recent studies, there were about 2,600 citizens living in Naustdal in 2019 (IMDi 2019). There are some lakes located in the upper catchments and in the lower basin. Because of its position and inflows from different sub-catchments one lake was handled as an olake (see sub-catchment 6 in Figure 8).

According to the land use (SLC) a small part of Nausta is glaciated area. The area classified as glacier has a size of about 10,000 m^2. Using an area-volume relation proposed by Bahr et al. (1997) and

Bahr et al. (2015), the glacier volume is estimated with 57,164 m^3. Assuming that the glacier is fully melting in summer (April-September) leads to a mean discharge of 0.004 m^3/s. Even if the glacier volume would melt in only one month, there would just be a discharge of 0.02 m^3/s. Since these discharges are low compared to measured outflows out of the catchment, a glacier influence was not considered in the model. Another reason for not including glacier influence is that according to another source there is no glacier located in Nausta directly (NVE Atlas). Although there are seven glaciers near, all of them are located north of the mountains outside the catchment (see Figure 48 in Appendix B.1). A map of the SLCs is illustrated in Figure 9. A summary of characteristics of Nausta can be found in Table 1.

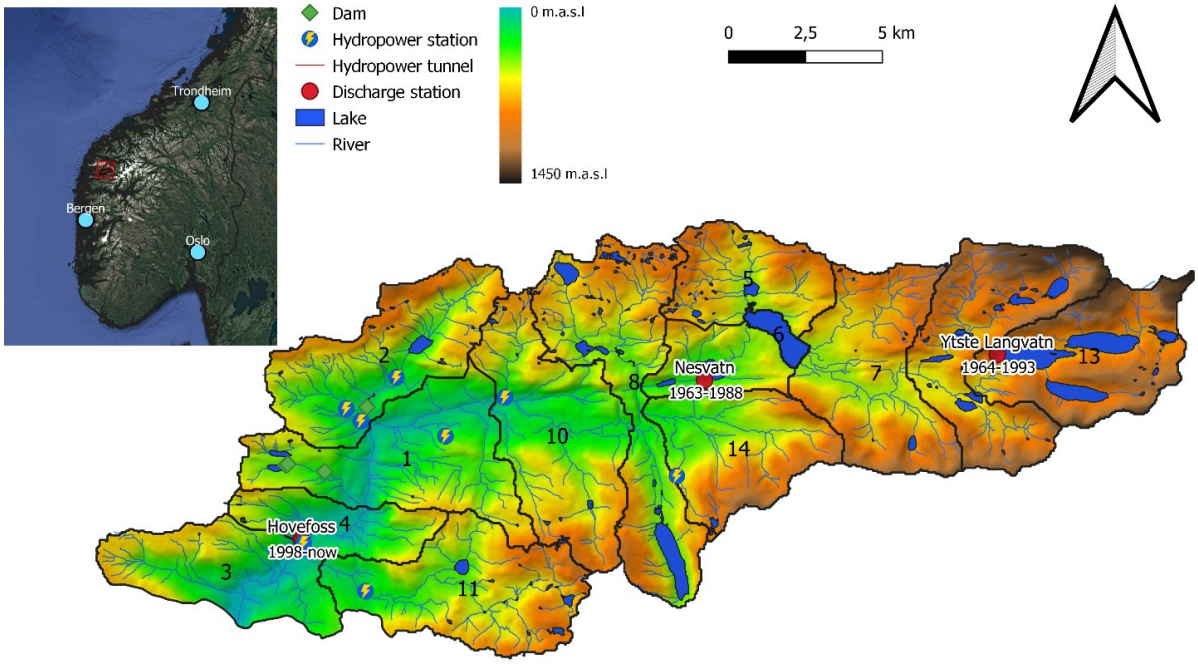


Figure 8: Location, elevation, rivers, lakes and discharge measurement stations of Nausta

The catchment includes three discharge measurement stations, that are in the upper, middle and lower basin. Since Yveste Langvatn and Nesvatn station weren't collecting data for more than 20 years, they were not included in this study. The station used for model calibration is named Hovefoss and is collecting data since 1998 with a timestep of 30 minutes (see also Table 5 in chapter 3.3.3).

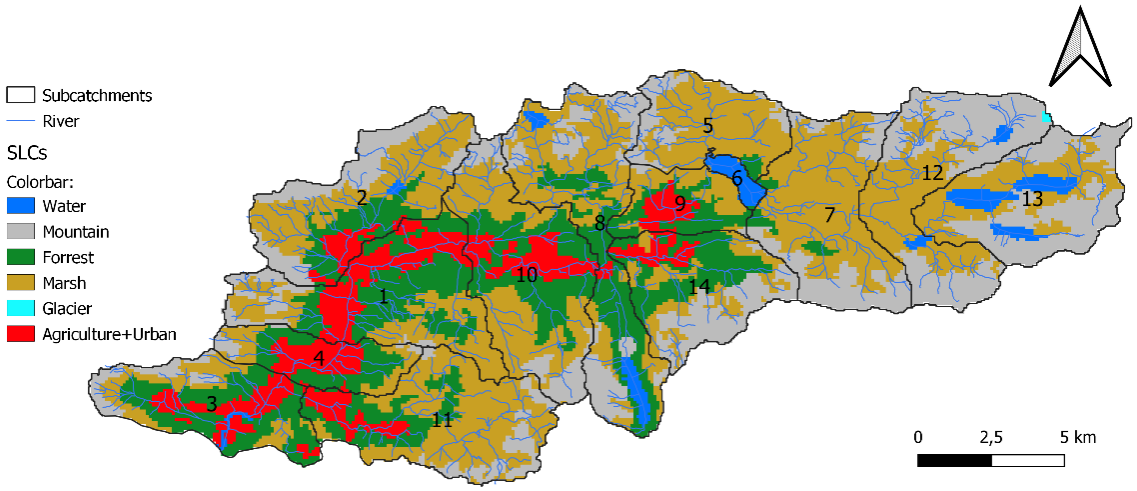


Figure 9: SLCs of Nausta

As seen in Figure 10, there is a high seasonal variability in the catchment. Beginning in September, the temperature in the area is dropping and reaching a mean daily temperature below 0 °C in November. The discharge in September is high, because of a high amount of precipitation in autumn, but decreases together with temperature due to more solid precipitation and frost in winter. The lowest temperatures are in the winter months of January and February with daily mean temperatures of about -5 °C. Discharge reaches its minimum in February and March. When temperatures rise above 0 °C in April, the discharge increases fast due to snow melting in the catchment. The discharge peak is in Mai with about 30 m^3/s, although there is not a lot of precipitation in that month. In summer discharges are decreasing caused by less precipitation and high evapotranspiration.

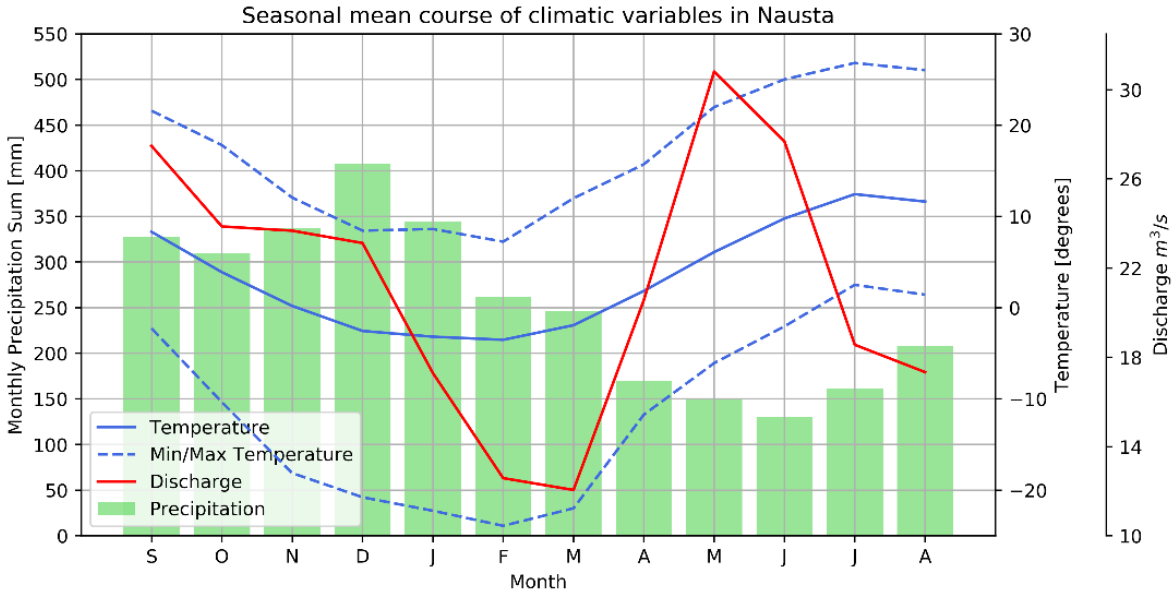


Figure 10: Seasonal mean course of temperature, precipitation and discharge (measured in Hovefoss) in a time span of 20 years (2000-2020). Data: SeNorge2018

3.2.2 Gaula

Gaula is with about 3,600 km^2 the largest chosen catchment. It consists of over 200 sub-catchments and includes eleven discharge measurement stations. It was chosen because of its high amount of available discharge data with long time series as well as a large area, where the distribution of precipitation has a higher influence in calculating the runoff. Gaula is located south of Trondheim with a river mouth west of the city. Elevation reaches from 0 to 1,319 m a.s.l. with a mean slope of 8 °. Some lakes are located in the upper basin as well as four large lakes north-east of Gaula. These lakes are reservoirs for hydropower stations. Because of hydropower paths between reservoirs and hydropower stations, the flow is not natural everywhere. These catchments with unnatural flowing conditions were not included for the calibration. There are also smaller hydropower stations west of the basin. It was assumed, that they don't impact the hydrology, since they are not using reservoirs.

In higher elevated areas of Gaula, the land characteristics are mostly marsh and mountain areas. In lower elevated areas are forests, agriculture and urban areas (see Figure 50 in Appendix B.1 and Table 1). The city Melhus is located in the lower basin with a population of 17,000 citizens (Statistics Norway 2021). In addition, parts of Trondheim are included in northern parts of the catchment.

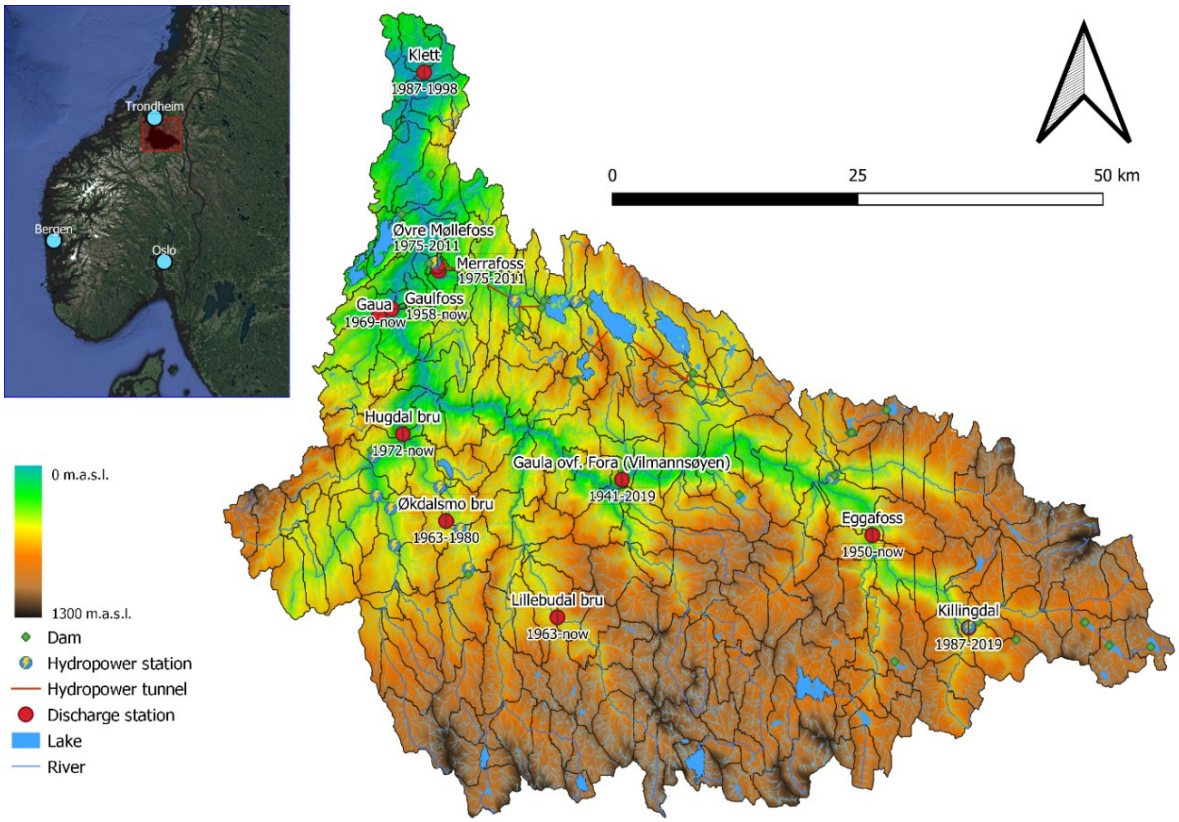


Figure 11: Location, elevation, rivers, lakes and discharge measurement stations of Gaula

There are seven discharge measurement stations located in the area. Five of them are currently active, with two additional stations collecting data until 2019. All of these have a timestep of one hour or less since 2000 and are therefore suitable for calibration and validation of the model. Figure 12 visualizes the upstream area for each gauge. The characteristics of these catchments are summarized in Table 3.

Table 3: Characteristics of Sub-catchments in Gaula

Catchment	SUBID	River	Upstream Area [km^2]	No. Upstream basins	Mean Elevation [m]	Mean Slope [°]
Gaua	200	Gaua	80	4	465	9.8
Lillebudal bru	242	Bua	171	9	913	9.5
Killingdal*	52	Gaula	226	11	882	7.1
Hugdal bru	102	Sokna	545	33	653	8.5
Eggafoss	126	Gaula	655	38	833	7.6
Gaula ovf. Fora*	124	Gaula	1404	84	746	7.6
Gaulfoss	108	Gaula	3060	193	730	7.7

* only/mostly daily discharge series available

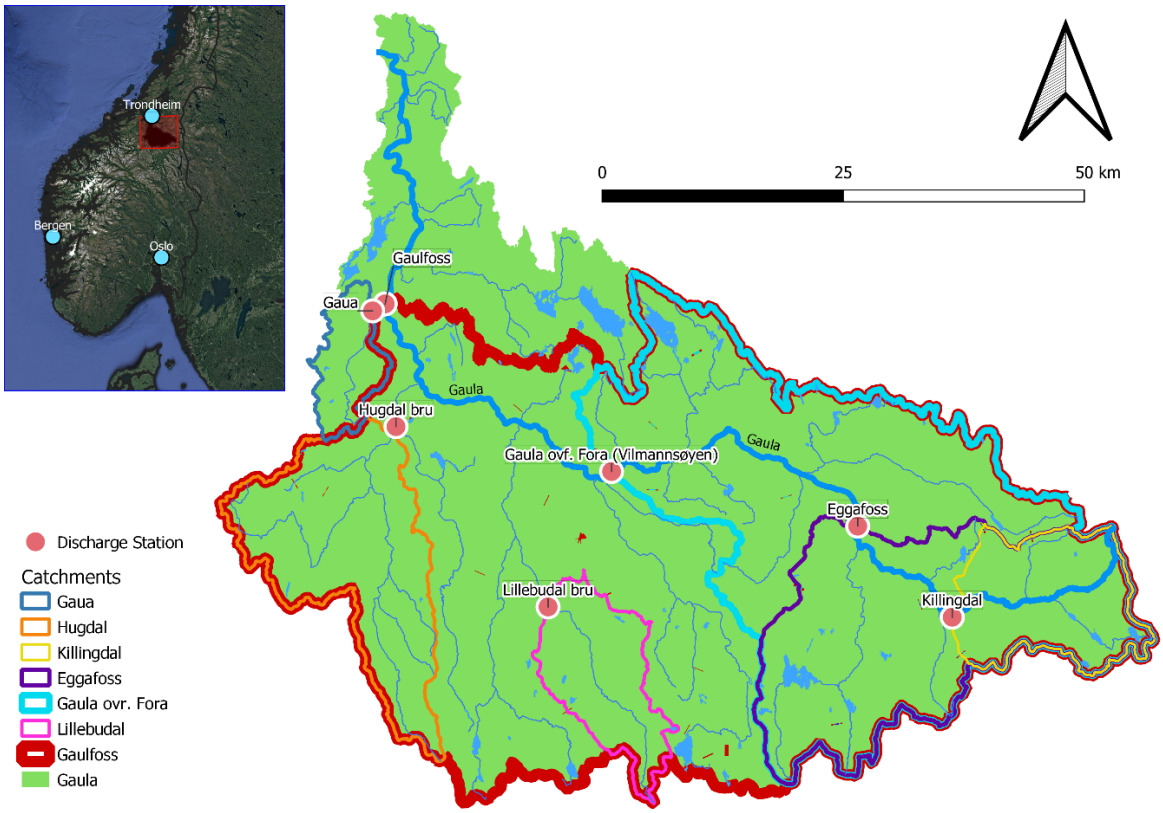


Figure 12: Upstream area for each of the gauges located in Gaula

The seasonal variability is visualized in Figure 51 (Appendix B.1). In agreement to Gaulas hydrological regime (Gottschalk et al. 1979), the discharge is dominated by snow melt in the spring time and has low flow conditions in winter. The mean temperature of Gaula is negative between November and April, with a minimum temperature below -30 °C. There are no strong variabilities in the precipitation except for a small decrease of precipitation in spring.

3.2.3 Usma

Usma is located south of Trondheim in Trøndelag and is with an area of about 42 km^2 the smallest of the study sites. It was chosen because of a lack of weather stations near the catchment. The nearest rain gauge is located 30 km away from Usma and is likely to not represent meteorological situations of the catchment correctly. This evaluation is part of the HydroCen project (https://www.ntnu.no/hydrocen/5.-vannkraftressurser-og-flerbruksinteresser), where more detailed hydrological data all over Norway will be provided.

The Usma river is an inlet into Nidelva, a large river that ends in Trondheim. Usma is mostly unregulated, with a dam located at the outlet of the basin. There, water is diverted through a hydropower tunnel into a hydropower station with a capacity of 10 MW. There is one lake with an area of 0.4 km^2 in Usma. Because of the size and four upstream sub-catchments connected with the lake, it was modelled as a sub-catchment (*olake*). Usma has an elevation between 500 and 1,050 m a.s.l with a mean slope of 10.6 °. The area consists mostly of marsh, mountain and forest areas, with no urban or agriculture land use. Table 1 summarizes the characteristics of Usma.

There is one discharge measurement station available. This station is maintained by TrønderEnergi AS for observing the hydropower station. There is no information about the quality of this measurement. An overview over the catchment is given in Figure 13 and a map of land uses can be found in Appendix B.1, Figure 52.

Looking at climatic parameters of Usma (see Figure 48, Appendix B.1), precipitation, discharge and temperature share the same variability as Gaula, due to the close location of these catchments. The discharge agrees with mountain regimes defined by Gottschalk et al. (1979).

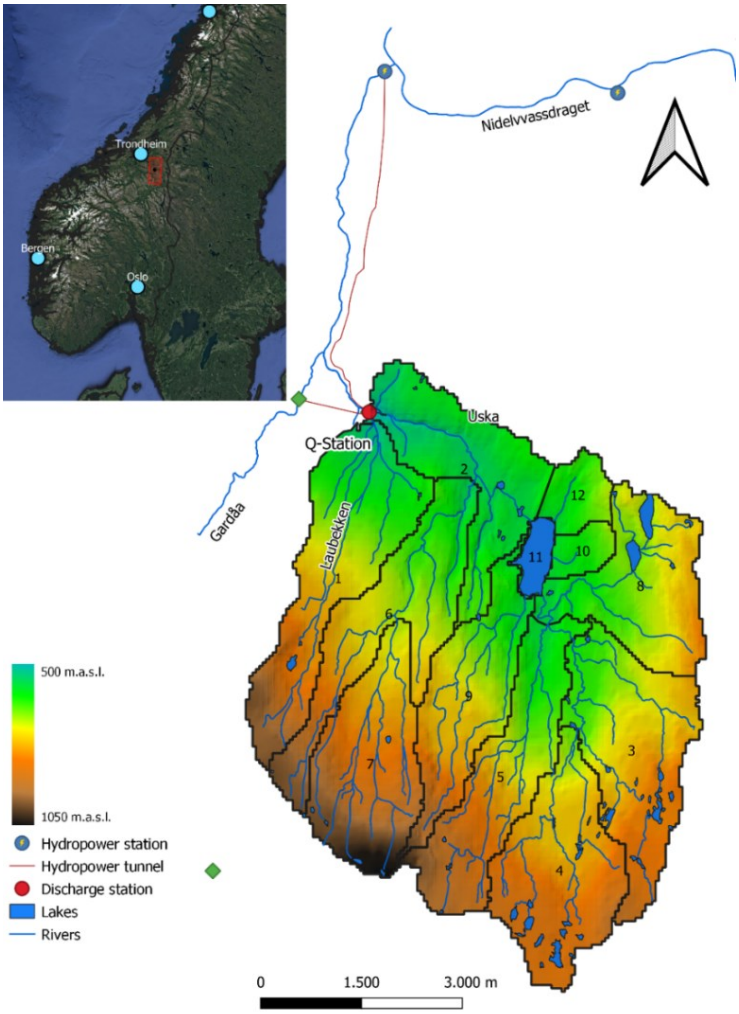


Figure 13: Location, elevation, rivers, lakes and hydropower locations of the Usma catchment

3.2.4 Surna

Surna is located in Trøndelag with a river mouth at the sea. The catchment has a high elevation in the south and is flatter in the north. The area is 1,221 km^2 with a mean elevation of 648 m a.s.l. Two large lakes are located in the basin that are influenced by glacier melt. Surna is highly regulated by power plants, especially through the lakes in the catchment that are used as reservoirs. Because of a decrease of discharges during low flow conditions in Surna, there has been a decrease of fish population. Especially in winter, the fish can't adapt to low flow conditions (Thorstad et al. 2003). To analyse the severity of the intervention into nature, the population of fish will be measured and analysed during the DynaVann project (https://www.ntnu.no/hydrocen/dynavann). To compare fish population with discharges (especially low flows), a hydrological model is created that generates discharge data for further analysis.

Not regulated areas of the catchment are upper parts of the basin. The only available discharge station in a not regulated area is Rinna (see red borders in Figure 14). The model will therefore be calibrated for Rinna and the resulting parameters will be applied on the rest of the catchment. The

discharge data is published by the Norwegian Water Resources and Energy Directorate (NVE). The gauge is located upstream of a regulated reservoir and not influenced by it.
The area in Surna in the north is primary classified as mountainous and marsh. In the south (lower elevation) the SLCs are mostly forest and agriculture. SLCs are visualized in Figure 54, Appendix B.1. When looking at climatic parameters of Rinna (see also Figure 55, Appendix B.1), the mountainous regime characteristics can be seen. The mean temperature is below 0 °C between November and April. Discharge is dominated by snow melt beginning in April but reaching its peak in Mai and June. The precipitation amount is highest in autumn and winter, even so the discharge is the lowest during this time.
There are three glaciers located inside the Rinna catchment. They each have an area of about 0.04-0.05 km^2 (NVE Atlas). By using the relation between area and surface suggested in Bahr et al. (2015), a volume of 0.0015 km^3 (1,530,530 m^3) of glaciated area can be assumed. The annual discharge of glacier would then be (assuming a constant outflow and a total melt of discharge within a year) 0.05 m^3/s. If all glaciers would melt within a month, the discharge would be 0.6 m^3/s. In the month with the lowest discharges and mean temperatures above 0 °C (October) that would be 17 % of the mean discharge. Nevertheless, glaciers where not implemented in the model, because they would add more parameters and uncertainty and probably are not influenced that much by the precipitation input.

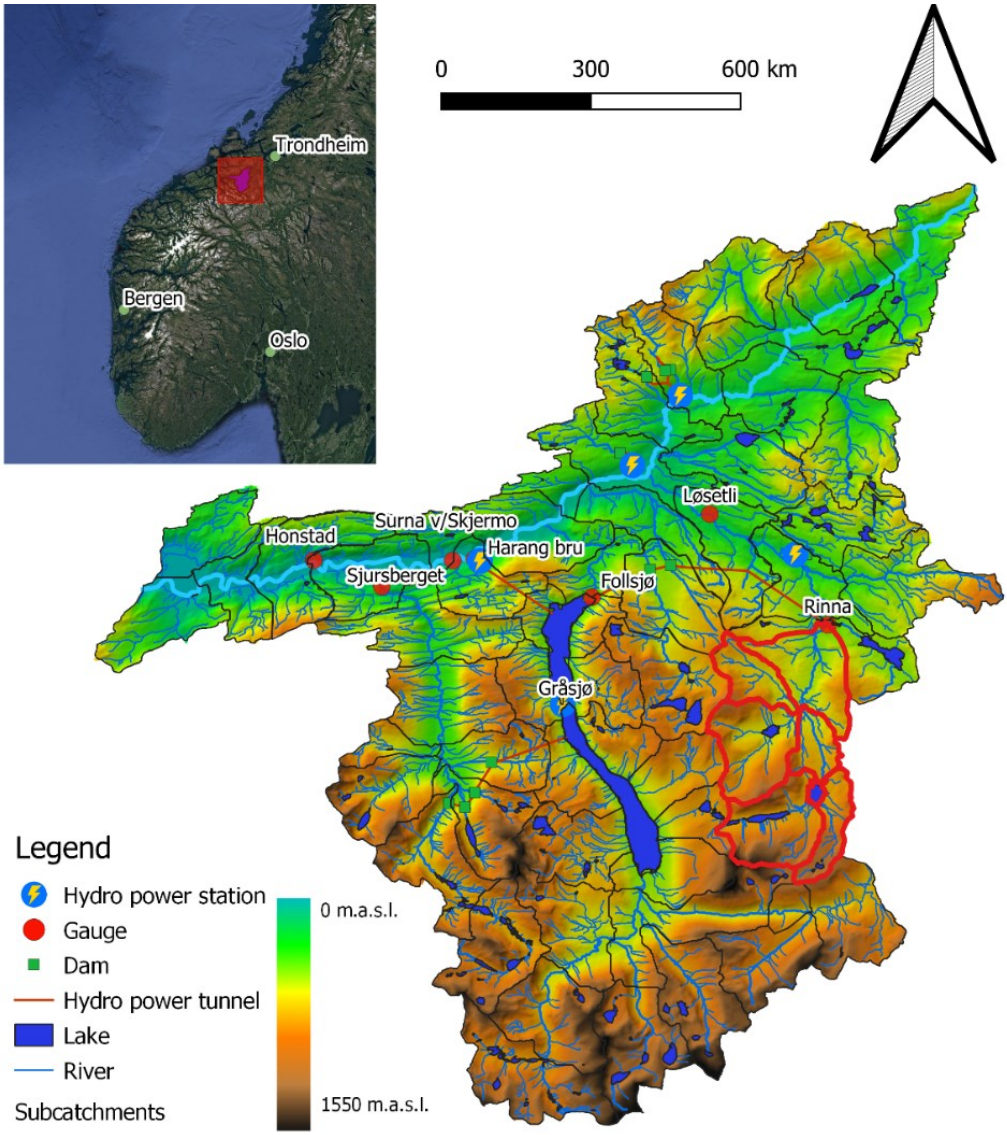


Figure 14: Location, elevation, rivers, lakes and hydropower locations of the Surna catchment

3.2.5 Grunnåi

The Grunnåi river is located in Telemark, Norway. The catchment has an area of about 85 km^2 and a mean elevation of 960 m a.s.l. According to Gottschalk et al. (1979) the catchment is located in the inland regime. Relating to the TwinLab project (https://www.ntnu.no/hydrocen/twinlab), the aim is to gain a better understanding of low flow conditions and to couple hydrological, hydraulic and other models that are connected to a hydropower plant.

There is one hydropower plant located in the catchment, with a small reservoir above the hydro tunnel. This power station is not considered in the hydrological model, because the natural state of the basin is going to be regarded. Two discharge stations are located upstream of the hydro power station. The data is provided by Skagerak Kraft AS. One of the gauges only has data before 2000 and was not considered in this thesis. The other gauge is the calculated inlet into the reservoir between 2010 and 2022 and was used for the calibration.

There is no city located in the catchment. In the lower elevated areas, the land use is mostly forest (32 %), in the higher elevated areas the land class is mountainous (24 %) and marsh (41 %). Although there are many lakes located in the basin, only the reservoir was considered as an *olake*. A small area is agriculture/city where some houses/cabins are located. But they are widespread and will not have a strong influence for the hydrology.

The temperatures in the catchment are in general low with a mean temperature above 0 °C only between April and October. The precipitation is the lowest in winter and spring and the highest in autumn. Caused by snow melt, the discharge reaches a peak in Mai, followed by discharge dominated by rain and higher discharges in autumn and low discharges in winter because of snow accumulation. This is also in good agreement to the hydrological inland regime (Gottschalk et al. 1979). A modelling challenge in this catchment are the small lakes in it. They are not considered as water bodies in the land use classification, but probably have a dampening effect on the discharge. That's why they were included manually into the SLCs by calculating the lake area of each sub-catchment and including them as *ilakes*.

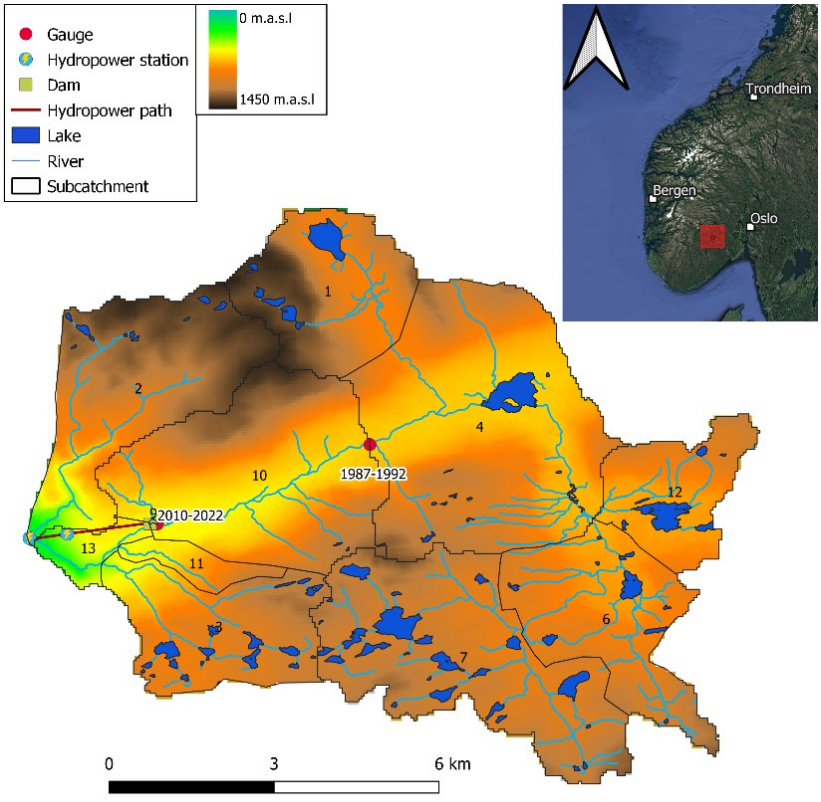


Figure 15: Location, elevation, rivers, lakes and hydropower locations of the Grunnåi catchment

3.3 Data

This section is a summary of precipitation products in Norway. Here, Norwegian radars and radar correction, spatial interpolated gauge data and a numerical model-based product are described. Additionally, evapotranspiration, discharge data and the data availability are included.

3.3.1 Norwegian Radars

"To put a weather radar in a mountainous region is like pitching a tent in a snowstorm: The practical use is obvious and large — but so are the problems." (Germann and Joss). This cite describes the difficulties when trying to measure radar data in mountainous areas like the Norwegian inland. Beam clutter, attenuation and beam blockage are the main error sources here. In a study from Abdella et al. (2012), beam blockage of radar stations in Norway was detected. The results showed a high amount of beam blockage in the direction of the inland of most of the radar stations.
There are currently ten radar gauges used in Norway, most of them located at the coast (see Figure 16). They cover all of the land area with a radius of 240 km (Bye 2013). Radar data is measured with a sampling rate of 7.5 minutes (Lussana et al. 2016a), but is accumulated hourly before being published online. The frequency of these radars is C-Band and most of them are horizontal polarized. Two radar stations (Berlevåg and Hurum) are dual polarized. Gridded, hourly radar precipitation (in mm/h) is available since 2010, radar reflectivity with clutter and blockage information as well as precipitation classifications are available since 2018.
Before being published by the Norwegian Meteorological Institute (MET), radar data is going through extensive quality control and transformation done by a ProRad system (Elo 2012). Each radar measures the reflectivity for up to 12 elevations (between 0.5 and 15 degrees), with a 360 degrees azimuth. Each elevation is then corrected with the following filters (Elo 2012):

- Frequency of Occurrence: determination of time dependant clutter
- Doppler filter: determination of ground clutter (radial velocity close to zero)
- Fast Marching Method: determination of sea clutter
- Determination of sun flare: calculating the total variation of a line segment
- Classification of precipitation type: using temperature and humidity data to separate snow, sleet and rain

After filtering the raw data, the gaps are reconstructed (Sivasubramaniam et al. 2018). To generate the Surface Rainfall Intensity, the Vertical Profile of Reflectivity is identified and corrected to diminish the bright band effect and increase lower rainfall rates at longer ranges for stratiform and convective events (Elo 2012). This approach takes the vertical variability into account and refers radar measurements to reference level. That is important, because the height of the radar beam increases with range and cannot be compared to the ground truth in higher ranges (Elo 2012).
To calculate the precipitation rate from the radar reflectivity Marshall and Palmer (1948) (Eq. 25) is applied, not considering the changes of the Z-R relation caused by different precipitation types (e.g. snow and rain). That's why Sivasubramaniam et al. (2018) suggested to use air temperatures for calculating the precipitation from radar reflectivity. The temperature yields information about the precipitation type and creates a variable Z-R relation. This resulted in an improvement of rain estimation of more than 80 % compared with gauge data, with the highest improvements with temperatures below 10 °C. In a further study Sivasubramaniam et al. (2019) combined radar precipitation data with interpolated gauge precipitation. The aim of this study was to improve the spatiotemporal resolution of precipitation estimates. They concluded that radar data underestimates the precipitation compared to gauge observations, even after merging them. Nevertheless, this approach created better results of radar data.

In general, the amount of radar stations in Norway is not enough to create a land covering observation field. The reason for that are the mountains challenging the observations. That's why radar data should be best used for catchments right next to the station or somewhere without any shielding effects (Lusanna 2022, personal communication, 10th March). An overview of radar location is visualized in Figure 16, technical details of each of the radar stations are summarized in Table 4.

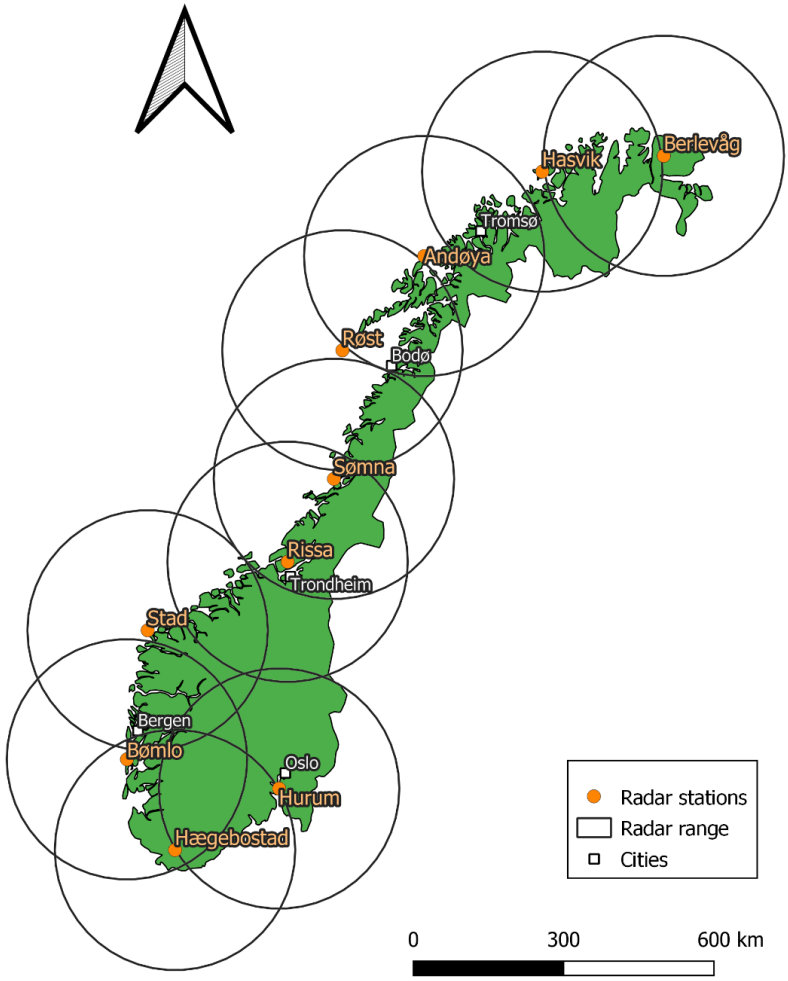


Figure 16: Overview over radar coverage of Norway

Table 4: Summary of available weather radars in Norway, information taken from Bye (2013)

Name	Height-Station [m]	Height-Antenna [m]	Data since	Frequency	Polarization
Rissa	605	611	01.01.2003	C-Band (5,640 GHz)	Horizontal
Andøya	433	436	01.01.2007	C-Band (5,620 GHz)	Horizontal
Berlevåg	455	474	01.10.2012	C-Band (5,640 GHz)	Horizontal+Vertical
Bømlo	95	104	01.01.2002	C-Band (5,640 GHz)	Horizontal
Hægebostad	622	631	01.01.2000	C-Band (5,640 GHz)	Horizontal
Hasvik	437	444	01.09.2008	C-Band (5,640 GHz)	Horizontal
Hurum	347	363	01.11.2010	C-Band (5,620 GHz)	Horizontal+Vertical
Røst	3	17	01.01.2004	C-Band (5,640 GHz)	Horizontal
Sømna	315	10	2014	C-Band (5,640 GHz)	Horizontal
Stad	496	508	01.10.2009	C-Band (5,640 GHz)	Horizontal

3.3.2 Norwegian NWP

In a cooperation between the Swedish SMHI and the Norwegian MET, called Meteorological Cooperation on Operational Numerical Weather Prediction (MetCoOp), the regional weather prediction system AROME-MetCoOp was developed for operational weather forecast in Nordic countries (Müller et al. 2017). This model has a resolution of 2.5 km. To adapt it to the Nordic region, it was calibrated with radar reflectivity data, ground-based satellite observations and gauge observations (Müller et al. 2017). For convective prediction, the MetCoOp Ensemble Prediction system (MEPS) is operated in Finland, Norway and Sweden. It is based on AROME. MEPS runs four times a day and has ten ensemble members. It adds more value to precipitation events in summer. A MEPS-Arctic version was developed as well, that covers Svalbard and the surrounding areas. Figure 17 shows the area covered of numerical models.

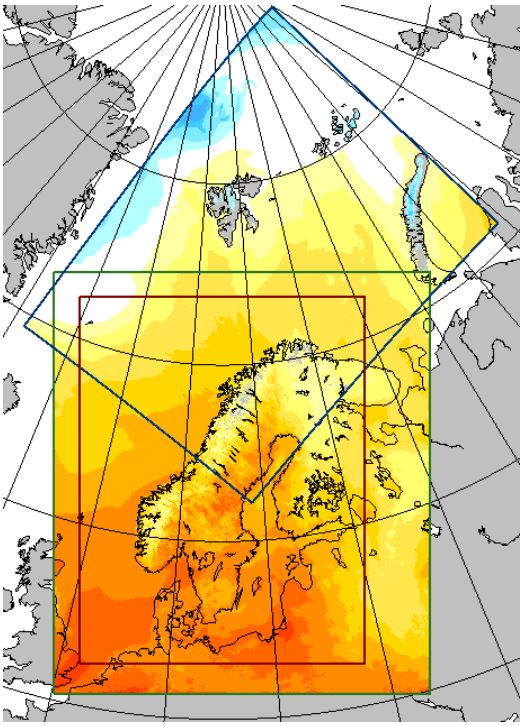


Figure 17: Model field of MEPS-Nordic and MEPS-Arctic (Nipen 2022)

3.3.3 Temperature and precipitation products published by MET

In this chapter, available gridded meteorological datasets over Norway are described. All of them were published by MET and are publicly available.

seNorge2

SeNorge2 is the second version of spatial interpolated temperature and precipitation data. Hourly data is available between 2010 to 2017 and covers all of Norway in a 1 km x 1 km grid. The product was calculated by using weather and climate stations in a scale separation approach (Lussana et al. 2016b). Temperature data is interpolated by first creating a pseudo background using a Bayesian method on a larger scale to assume effects of atmospheric dynamics and then using optimal interpolation with that pseudo background and more observation points on a smaller scale. Temperature is assumed to be at a height of 2 m. Precipitation is also calculated by using spatial scale separation. Here, geographical coordinates (latitude, longitude) and elevation were included in an optimal interpolation (Lussana et al. 2018). Observation data is not corrected from wind-induced under-catch.

There is no study published about the hourly precipitation field. Nevertheless, the daily seNorge2 product was described and evaluated in several studies. In Lussana et al. (2017) the product was evaluated by checking the runoff-precipitation balance, testing a snow model and using the data as input in an hydrological model. They concluded that the station network is too sparse to detect small scale precipitation events like thunderstorms. The product tends to underestimate the precipitation amount. Snow is especially underestimated in northern Norway, rain at the west coast and in the mountains (Lussana et al. 2018). Intense precipitation is more underestimated than weak precipitation. Since hourly precipitation is even more locally variable and there are less gauges available measuring hourly precipitation, these results are likely to appear as well or even more in the hourly product. Because of the lack of gauges and the uncertainty of hourly gridded data, the hourly seNorge2 product is not released anymore.

seNorge2018

SeNorge2018 is the newest version of the seNorge products. It was developed by MET and includes precipitation and temperature data from 1957 to 2017 in a daily timestep (Lussana et al. 2019b) which covers all of Norway in a 1 km x 1 km grid. To spatially interpolate precipitation, a climate model, observation data and a wind model to calculate wind-induced under catch of gauges was applied. For bias correction, monthly precipitation sums of gauge data are used. According to Lussana et al. (2019b) seNorge2018 varies highly from seNorge2, but since the precipitation product of seNorge2018 is not available in hourly timesteps, it could not be used in the model directly.

Temperature interpolation is done by a spatial scale-separation approach, starting with larger fields and using predictions to adjust it and then go into a smaller scale and use local observations, depending on the density of gauges. A digital elevation map (DEM) is also included in the interpolation. Mean, maximum and minimum temperatures are calculated separately and then checked for physical mistakes. Having only a small amount of bias, the seNorge2018 temperature has the highest bias for minimum temperatures in winter (Lussana et al. 2019b). SeNorge2018 was used for calculating the seasonality of the catchments (because a long time series is available) as well as for calculating the potential evapotranspiration (see chapter 4.4).

MEPS Nordic

The MEPS Nordic product (referred to as NWP) is published by MET Norway. It includes hourly precipitation and temperature data since 09/2013 and covers all of Norway in a 2.5 km x 2.5 km grid. There are more parameters included like relative humidity, wind speed and direction, radiation, air pressure, surface albedo and cloud coverage. This product combines numerical model data with observations. The interpolation used for that is an Bayesian interpolation, similar to the one described in Uboldi et al. (2008). It combines the numerical model as the background with observations as the ground truth. The numerical model used as background is a deterministic version of MEPS Nordic. It is corrected by a scaling factor and localization procedure (Lussana et al. 2019a). The observations used for the interpolation are classified according to their representation, which can vary over time (Lussana et al. 2019a).

For the temperature station data is used as observations. To improve the gauge density, quality controlled personal gauges were included in that process (Nipen et al. 2020). According to Lussana et al. (2019a) the method created more accurate results. Observational precipitation data is a combination of personal and public gauge stations and post-processed radar data. According to Lussana, the radar data is the biggest uncertainty (Lusanna 2022, personal communication, 10th March). An example of the combination of model data with observations is visualized in Figure 18. Here, the model used as a background was a non-deterministic version of MEPS and therefore has an uncertainty. The figure shows the adjustment of the background to the observations.

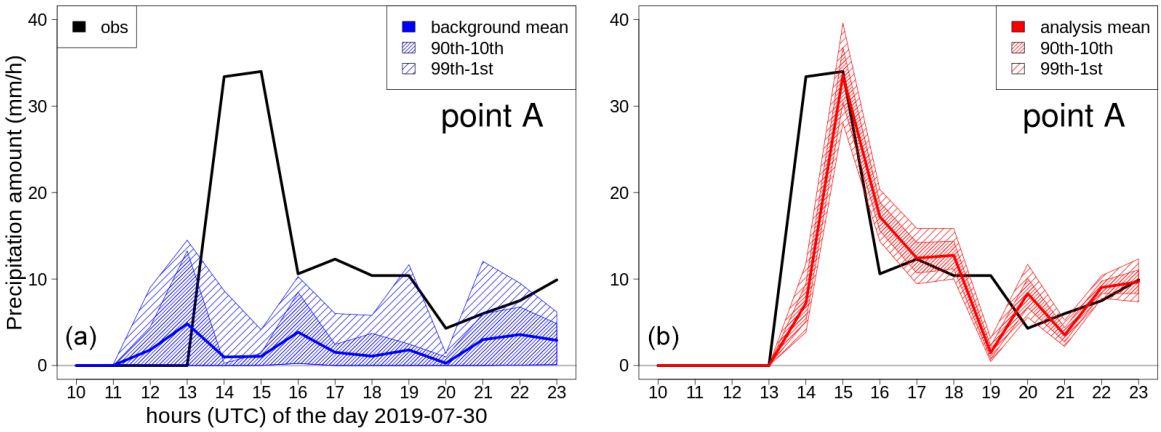


Figure 18: Example of the interpolation method used for NWP: Precipitation over time at an example point A. On the left is the raw background (blue) and observations (black). On the right sides the interpolated product (red). Source: Lussana et al. 2021

NWP was not yet evaluated in its final version, but in Frogner et al. (2019) MEPS-Nordic (background of NWP) was evaluated. They evaluated 54 precipitation events, especially small-scale heavy precipitation events. They compared the forecast versus observations. They conclude that MEPS adds value to small scale predictions, because of the uncertainty addition. Especially in summer when predictability is lower. Still, small scale precipitation has a limited predictability.
Another comparable approach is described in Lussana et al. (2021) where a numerical model was used as the background of an optimal interpolation as well. It differs from the operational NWP product, because a non-deterministic MEPS version was used. That's why the results cannot be directly compared to NWP. Still, the study showed that including observational data to the model added value over the background and improved the precipitation estimate field.

KliNoGrid RR-Rad

The most radar dependent product analysed in this thesis is called Norwegian Climate Gridded dataset (KliNoGrid) and is a combination of radar and gauge data. It was developed at MET in 2016 to be used for snow and hydrological modelling. Radar and gauge data was combined by using an optimal interpolation between spatial information of radar measurements and measured precipitation at gauge locations. The dataset is available for hourly and daily timesteps, but known to have a better performance for daily timesteps caused by smaller spatial variability of daily precipitation as well as a denser network of gauges measuring daily precipitation (Lussana et al. 2016a). To calculate the rain rate from radar reflectivity Marshall and Palmer (see Eq. 25) was used for summer and winter. KliNoGrid showed better results for summer than for winter (Lussana et al. 2016a), which may be caused by different relations between radar reflectivity and precipitation rate for snow and rain (Fujiyoshi et al. 1990). To improve the quality, only trustworthy gauges were included in the interpolation. Attenuation was corrected by using long-term accumulated precipitation fields from radar data (Lussana et al. 2016a).

3.3.4 Discharge

NVE operates a gauging system all over Norway measuring the discharge in rivers and lakes and publishes them online. Available measurements are water flow (Q), water level (W), water temperature (T_W) and air temperature (T_A). The discharge is calculated by using a rating curve including the measured water level (NVE 2015c). The rating curve is generated by measuring several discharges of the river during altering water levels. The discharge can be measured by Acoustic Doppler Current Profilers, tracers or flow meters (NVE 2015a). A measurement error that has to be

taken into account in Norway are frozen lakes or rivers, that change the water level and create inaccurate discharge calculations (NVE 2015b). The data is quality controlled by NVE. There are three different quality classes: no quality control (1), primarily quality control (2) and secondary quality control (3).
Some timeseries have a long measurement period others were only installed in the last decades. Timesteps vary between daily measurements and 15 minutes. Table 5 gives a summary over available discharge stations at study sites in the given time interval from 2010 to 2020.

Table 5: Summary of available discharge measurement stations in the catchments

Catchment	Name	Station ID	Measured Parameters	Availability	Timestep	Quality
Nausta	Hovefoss	84.11.0	Q, W, T_W, T_A	1998-now	30 min	2
Usma	TrønderEnergi AS	-	Q	12/2013-2017	1 h	-
Grunnåi	Skagerak	-	Q	2010-2022	1 h	-
Surna	Rinna	122.8.0	Q, W, T_A	1969-now	1 h	2
Gaula	Gaula ovf. Fora	122.5.0	Q, T_W	1941-2019	1d	3 since 2019: 2
	Gaulfoss	122.9.0	Q, W, T_A	1958-now	30 min	2
	Eggafoss	122.11.0	Q, W, T_W, T_A	1950-now	30 min	2
	Lillebudal bru	122.14.0	Q, W, T_A	1963-now	since 1985: 1 h since 2007: 30 min	2
	Gaua	122.16.0	Q, W, T_W	1969-now	since 1995: 1 h	2
	Hugdal bru	122.17.0	Q, W, T_A	1972-now	since 1995: 30 min	2
	Killingdal	122.24.0	Q, W	1987-2019	1d	3 since 2019: 2

3.3.5 Evapotranspiration

As an independent observation to compare modelling results, MOD16 evapotranspiration was applied. MOD16 data is published by the Earth Observing System Data and Information System (EOSDIS). PET and AET are calculated using the equations established by Monteith (1965). The product combines Moderate Resolution Image Spectroradiometer (MODIS) data (land cover, leaf area index and albedo) with daily meteorological data from the Global Modelling and Assimilation Office (GMAO) (Mu et al. 2011). The dataset covers most of the vegetated area of the earth and is available with an eight-day timestep since 2000 with a resolution of 1 km. Uncertainties of MOD16 can be caused by uncertainties in the satellite images and meteorological data, measurement errors on site, differences of height between measurement and actual land surface and limitation in the calculation process of MOD16, because not every detail of physical processes can be taken into account (Anwar 2015).

Huang et al. (2019) compared the potential evapotranspiration of 42 stations in Norway with MOD16. They found out that MOD16 tends to overestimate potential evapotranspiration. A study from Anwar (2015), that compared MOD16 data with other evapotranspiration products in Peru also concluded that MOD16 tends to assume higher AET and PET values than other products.
Downloading the open-source data is possible using an online-tool, that extracts data by using a geodata file of the desired area for a specific time period. The output is a netcdf4 file that includes the spatial variability of AET and PET.

3.3.6 Data availability

Because this study used data from different sources, the available time series are not similar for all of them. The most important factor for the analysis is the available interval of meteorological datasets. SeNorge2 is available between 2010 and 2017. NWP between 09/2013 and 2022. To avoid accumulated snow at the beginning of the calibration or validation period, a hydrological year was defined to start in September. The two datasets overlap between 09/2013 and 2017. To have an initial period of one year (e.g. to fill soil water storages), the compared time slot was defined between 09/2014 and 09/2016. To get the best possible results in this interval, calibration was done at that time. The calibration period was chosen to be three years, validation two years. This leads to the calibration of seNorge2 being between 09/2013 and 09/2016 and NWP between 09/2014 and 09/2017.
In most of the catchments, discharge timeseries are available between 2010 to 2022. An exception is in Usma, where discharge data is only available between 12/2013 and 2017. Because of that, the hydrological model in Usma was not validated. Figure 19 shows a summary of the available time periods. Hourly temperature data is not available for the full time period as well. That's why the temperature is also a different input data.
In Table 17, Appendix B.2 the sources of data or information used in this thesis are summarized. Meteorological data was downloaded from a server of MET. Discharge data, as well as geodata of rivers, lakes and observation stations are mainly from NVE. Previous work done for HYPE in Norway is from Schönfelder (2017) and Schönfelder and Baclet (2022).

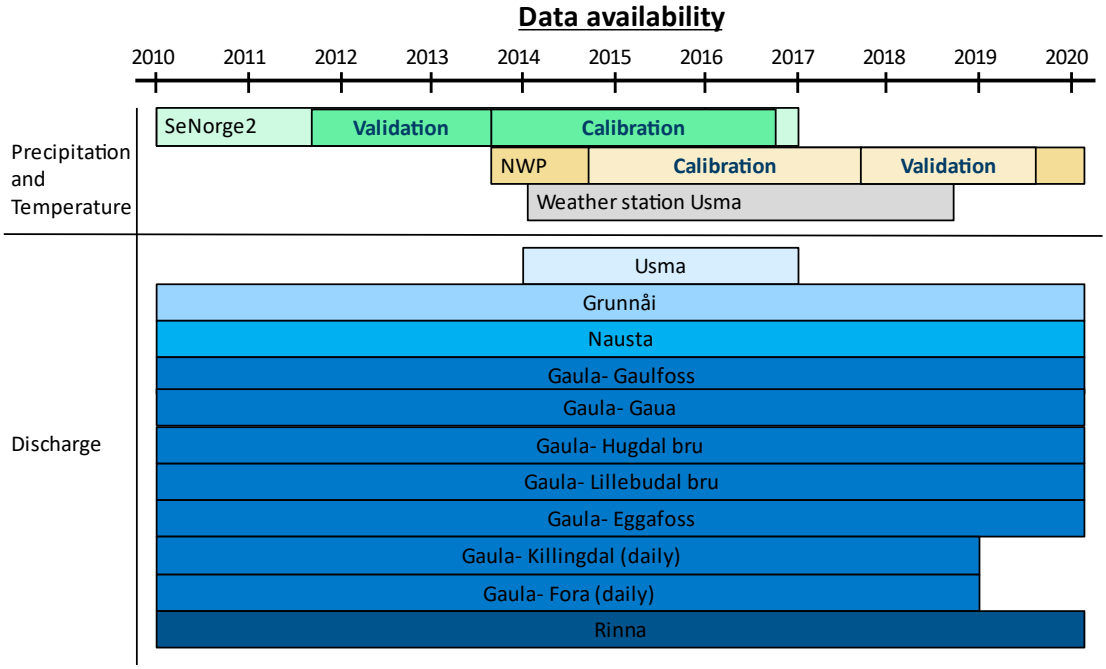


Figure 19: Data availability periods of the different observation products

4 Methods

This study was divided into two parts. First, the datasets were analysed and compared in general. For that, the annual precipitation sum was calculated and visualized over each catchment. To compare the datasets with each other, the Pearson and Spearman correlation of temperature and precipitation was calculated. Finally, a weather station owned by TrønderEnergi AS, which was not included in the calculation of the products, was compared with the grid point of the datasets to test their correlation. The second step of this study was modelling five catchments with the datasets using the HYPE model. For that, a calibration method was developed first. Each catchment was then calibrated with each dataset one time with precipitation correction and one time without. The results were analysed, and in a last step a sensitivity analysis was done for some of the catchments.

4.1 General evaluation of the precipitation and temperature products

In the first step of this analysis, precipitation products were compared with each other and one precipitation gauge. For that, annual precipitation sums, correlation between datasets and between a weather station was calculated.

In the first step, annual precipitation sums were calculated for seNorge2, NWP, radar precipitation and KliNoGridRad. The data is available as netcdf4 files on a download server and can be read directly using Python or R. A script was developed for reading netcdf4 files and summing up the precipitation for a specific time period. The precipitation sum was then again written into a netcdf4 file that could be visualised and analysed in GIS. The time period was chosen annual from September to September, because it was assumed that there will be no snow at the beginning of the period. The annual precipitation sum was visualized for 2014 (09/2014-09/2015) and 2015 (09/2015-09/2016), because this time period was also used for the comparison of the model results. SeNorge2 and NWP were further compared with each other. This evaluation provides information about the differences of the datasets and was further used to decide whether temperature data is comparable enough, so that it will not have a lot of influence on the model results, since it was not possible to use only one temperature product because of the different time periods (see chapter 3.3.4). For that reason, the Spearman correlation between seNorge2 and NWP was compared for each sub-catchment and lumped catchments. To create lumped temperature/precipitation data p_t the area A_i of each sub-catchment was multiplied with the temperature/precipitation $p_{t,i}$ per hour. This sum was then divided by the area of the whole catchment.

$$p_t = \frac{\sum_{i=1}^{n} p_{t,i} A_i}{\sum_{i=1}^{n} A_i} \qquad \textit{Eq. 26}$$

To further analyse the influence of snow, mixed precipitation and rain, the datasets were divided into temperatures -5 °C, above 5 °C and between those temperatures. For that, NWP temperature was the condition, creating a bias towards seNorge2 (especially the temperature correlation). Pearson correlation was considered as well, but Spearman was used at the end, because it is less effected by outliers.

The datasets were then compared with one ground observation. Publicly available gauges could not be used for that, because they are already included in the products. For that reason, a weather station maintained by TrønderEnergi AS was used as the ground truth. Temperature and precipitation data between 2014 and 2017 was obtained for that. The closest grid cell of the datasets was extracted out of the netcdf4 files. This was done for temperature and precipitation of seNorge2 and NWP. These timeseries were then compared to the station data by scattering them and calculating the correlation of seNorge2 and NWP. The bias was analysed by calculating the annual precipitation sum, mean temperature and divided in summer and winter.

Because weather station data was only available at one location, another data source was applied. According to Eq. 1, discharge data is the amount of precipitation minus evapotranspiration losses. This is only true when assuming there are no storage changes over a specific time interval. By assuming that there are no storage changes in the catchment from September to September of each year (because snow is already melted then), a mass balance was calculated for each year and each catchment. For that, discharge, evaporation and precipitation were calculated as volume (in mm, depending on the area of the catchment). AET is dependent on the potential evapotranspiration and the available water amount in the catchment. Since that information is difficult to estimate, AET from MOD16 was applied (see also chapter 4.2.4). The precipitation was than plotted against the sum of discharge and evapotranspiration. Assuming all observations are the truth and there are no storage changes over the time interval, these two values should be similar. The differences between precipitation and discharge with evapotranspiration can be called errors of the water balance of the catchment. These differences were compared for each basin.
An inverse modelling approach (chapter 2.1.3) was applied as well, to compare the precipitation calculated from discharge data and the precipitation products. This could not be done for snow events, because the simple model approach developed by Kirchner (2009) does not take snow accumulation and snow melt into consideration. First, the precipitation sum of the model and the datasets between June and October were compared. For that, a sensitivity curve was fitted for each catchment and the precipitation calculated from the resulting parameters.

4.2 Setting up and calibrating HYPE for the different catchments and input datasets

In this part, the catchments were calibrated with different input datasets. At first, the catchments were set up for HYPE by adjusting the sub-catchments and defining optimization boundaries. Afterwards, PET was calculated in an extra step because it was found that HYPE tends to compensate the water balance equation within the model by varying evapotranspiration rates. To create stable and reliable results, different calibration methods were tested and analysed in the next step. Here, the number of soil layers and their thickness was variated. Multi-gauge and low flow calibrations were further analysed. At the end of this section, the modelling process is summarized by listing the steps done for each study side.

4.2.1 Determination of sub-catchments and SLCs

The delineation and boundaries of sub-catchments were first implemented by Schönfelder (2017) and were further developed in the following years (Schönfelder and Baclet 2022). Initially, Schönfelder (2017) used a tool called WHIST (developed by SMHI) that calculates sub-catchments by using a DEM, flow accumulation and flow direction. In that process, forced points can be added to the calculation, which will result in sub-catchments with outlet points similar to the forcing point. That way, discharge gauges can be placed directly at the outlet point of a catchment to calibrate the model. In that process, also geographical data (e.g. mean elevation, mean slope) was calculated. Another output calculated in that process is the main water flow path. Schönfelder (2017) also included lake information into the determination of sub-catchments. For that, lakes were divided by area and lakes larger than 5 km^2 were considered as olakes, lakes smaller 5 km^2 as ilakes. In Schönfelder and Baclet (2022) a combination of Python code and Whitebox GAT was used for automatic routing. The SLCs are, in contrast to first model setups from Schönfelder (2017), based on the Coordination of information on the environment (CORINE) Land Cover (CLC) dataset which is published by the European Environment Agency (EEA 1993). This dataset consists of 44 classes, that were classified by using satellite images (Büttner et al. 2017). These classes were then

reclassified into five classes (water, marsh, mountain, forest, urban/agriculture) by Schönfelder and Baclet (2022).

Most of the available geographic data from previous work was used directly in this thesis. In some cases, sub-catchments had to be adjusted, e.g. because of separated local rivers, changing an ilake into an olake or discharge stations not located at the outlet of the catchment. In these cases, geographical data had to be recalculated. Figure 20 shows an example of manual changes done in Nausta. In this case one olake (sub-catchment 6) was added, because of two inflow points into the lake. Two sub-basins were divided to include a discharge measurement station (sub-basin 3 and 4) and to include inflows into the olake (sub-basin 5 and 9). Adjustments were done in Usma, Grunnåi and Surna (Rinna) as well.

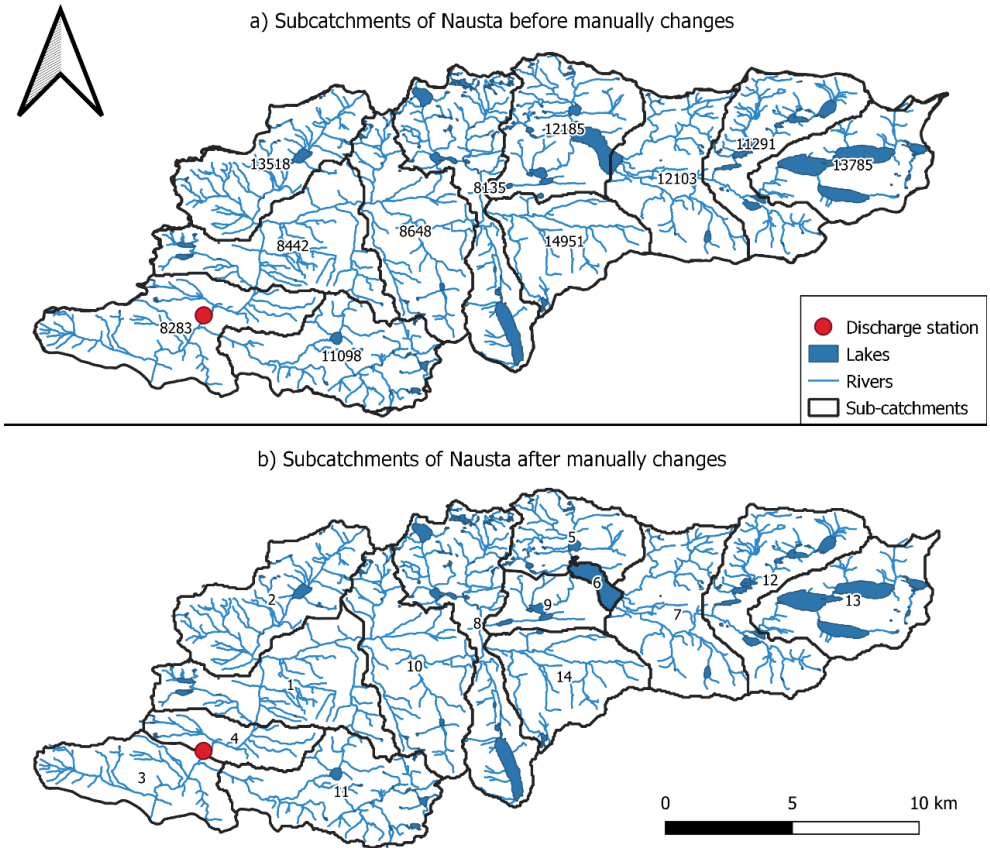


Figure 20: Nausta sub-catchments before manual changes (a) and after (b).

4.2.2 Parameter adjustment

Schönfelder and Baclet (2022) set up a set of parameters and parameter boundaries for HYPE for a daily timestep. When trying to change the model into an hourly timestep, the model showed a bad suitability. This was mainly caused by high evapotranspiration, that dried out the catchment causing basically no discharges in summer. The reason for the high evapotranspiration amount was the evapotranspiration parameter (*cevp*). Since this parameter usually is a constant for daily evapotranspiration it could not be adapted directly into the hourly model and was decreased (from 0.01-1 to 0-0.1). The same procedure was done with the snowmelt parameter *cmelt* (from 1-5 to 0-0.5).

According to Pers (2022b) all parameters with a unit of ts-1 are affected by changing the timestep. Additionally, *macrinf*, *surfmen* and *deepmem* are affected by changing the timestep. These parameters are less sensitive within the model and were not noted as an influence in the modelling

results at the beginning. The parameters were therefore not changed for the calibration of the catchments. This needs to be considered when analysing the results.
The number of parameters, that will be calibrated can be chosen by the user, as well as the parameter space. These calibrated parameters are summarized in Appendix A.2, Table 17. Most of them are similar to the ones used in Schönfelder and Baclet (2022). Some were changed so that the number of steps between the boundaries is between 100 and 150. The table also includes boundaries of each parameter and the number of steps between them.

4.2.3 Precipitation and temperature data extraction and interpolation

As mentioned in chapter 4.1 precipitation and temperature datasets are available publicly at the download server of MET. There, they are saved in nedcdf4 files that contain georeferenced gridded values. A R script developed by Schönfelder (2017) was applied to extract these values from the files for each sub-catchment by using a geodata file of the basin. This script was edited to open and read netcdf4 files directly online without downloading them first. This was done to save computer storage, since hourly datasets can take a lot of space. Another small change was done for the creating of the output text file. Originally, the script used to read all netcdf4 files first, before creating an output file. When working online, there is always a chance of losing internet connection or a failing server. That's why in the edited version, the results were written into a text file for each timestep. The scripts were applied for extracting seNorge2, NWP and seNorge2018 temperature and precipitation data.
The NWP product has a data gap between 2019-03-02 20:00 and 2019-03-04 03:00. Precipitation data between this time was set to zero, temperature was interpolated. When a sub-catchment was too small to fit into an index cell (this was especially the case with NWP, because it has a bigger grid than the other products) the index was defined manually to fit the closest grid cell of the sub-basin.
For two of the catchments, weather station data was included into the analysis. In Usma, a weather station located 30 km away from the catchment was evaluated to test whether it would be suitable for a hydrological model. For that, the measured precipitation of the weather station was applied uncorrected for each sub-catchment. Temperature data was not taken from the weather station, because it is more dependent on the elevation than precipitation. For that reason, NWP temperature data was mixed with weather station precipitation data.
In Grunnåi, daily precipitation data from MET (published at frost.met.no) was applied to analyse in the water balance. There are five gauges located around Grunnåi. To generate precipitation data for each sub-catchment (h_N), an Inverse Distance Weighting was applied. For that, the distance (d_i) between each gauge ($h_{N,i}$) to the centroid of each sub-catchment was calculated, and the weight of each gauge calculated according to Eq. 22. Not every gauge measured precipitation at each timestep. For that reason, the interpolation was done for each timestep, so that gauges with data gaps could be excluded from the calculation during that time. Figure 21 visualizes the location of the gauges and the sub-catchment centroids.

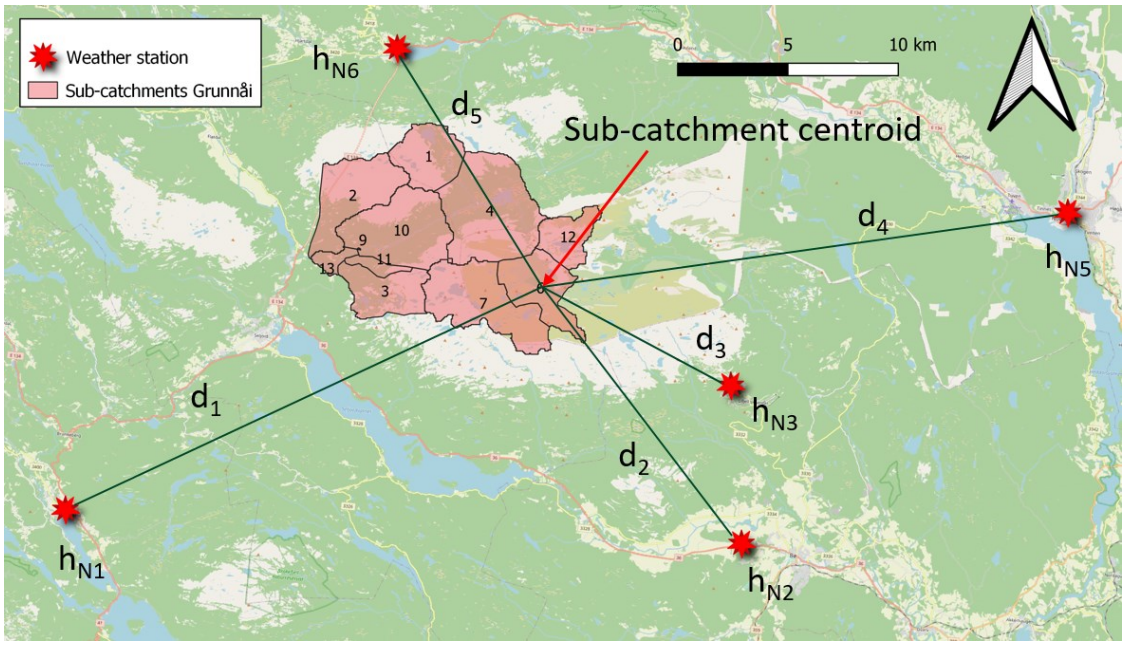


Figure 21: Visualization of Inverse-Distance-Weighting in Grunnåi with daily weather station data

4.2.4 Calculating potential evapotranspiration as an independent input in HYPE

Evapotranspiration is not a necessary input in HYPE. There are even different methods to calculate PET available as model options (e.g. Priestly-Taylor, FAO Penman-Monteith) (SMHI 2021). However, these methods can only be applied for daily timesteps, not for hourly data. The only available method for hourly timesteps is the default model. This method assumes, that PET (*epot*) is dependent on air temperature, with no evaporation below a threshold value (*ttmp*). Furthermore, PET is dependent on land use which is included with the rate parameter (*cevp*). To also account for seasonal changes, a parameter *cseason* is added to the equation. If necessary, a regional correction factor (*cevpcorr*) can be added.

$$epot = (cevp \times cseason) \times (T - ttmp) \times (1 + cevpcorr) \qquad \textit{Eq. 27}$$

Cevp and *ttmp* can be calibrated in an automatic calibration. Preliminary results showed that PET was not similar for each calibration. This may be caused by low sensitivity of the model towards changes of evapotranspiration. Additionally, a diurnal cycle is not considered, which means that on warm nights, the evapotranspiration was still high, although there was no radiation through sun available. For these reasons and to get a more physically and not calibration-based PET, it was calculated separately and added to HYPE as additional input data. For that purpose, the method of Hargreaves and Samani (1985) was applied. This is a simple approach using minimum (T_{min}) and maximum (T_{max}) temperature of a day to calculate the daily PET.

$$epot = 0.0023 \times R_A \times (T_{max} - T_{min})^{0.5} \times \left(\frac{(T_{max} - T_{min})}{2} + 17.8\right) \qquad \textit{Eq. 28}$$

Daily extra-terrestrial radiation (R_A) can be calculated according to FAO guidelines (Allan et al. 1998) by only using the latitude (*lat*) of the area and the day of the year (*J*). These two values are necessary to calculate latitude (φ) in rad, solar declination (δ) and sunset hour angle (ω_s).

$$R_A = \frac{24}{\pi} \times 4.92 \times (\omega_s \times sin(\varphi) + cos(\varphi) \times cos(\delta) \times sin(\omega_s)) \quad \textit{Eq. 29}$$

with $\varphi = \frac{\pi}{180} \times lat$ *Eq. 30*

$$\delta = 0.409 \times sin\left(\frac{2\pi}{365} \times J - 1.39\right) \quad \textit{Eq. 31}$$

$$\omega_s = arccos(-tan(\varphi) \times tan(\delta)) \quad \textit{Eq. 32}$$

According to studies from Hargreaves et al. (2003) and Zarei et al. (2015), the Hargreaves method shows acceptable results compared to Penman-Monteith, with less data input necessary.
In this thesis, evapotranspiration was calculated by using seNorge2018 daily minimum and maximum temperature data, because it was the most recent product that is available for the whole time period of this study. To transform daily evapotranspiration into hourly information, a weighting function was used, having the highest values in the middle of the day and no PET at night. In Figure 22 this weighting function is visualized for one day. The change of the diurnal cycle during the season was considered because the day and night time varies strongly between summer and winter in Norway. However, evapotranspiration was not considered to be a strong influence of the timing of the model and is more important for the general water balance instead of hourly values. That is why only one weighting function was applied for summer- and wintertime.

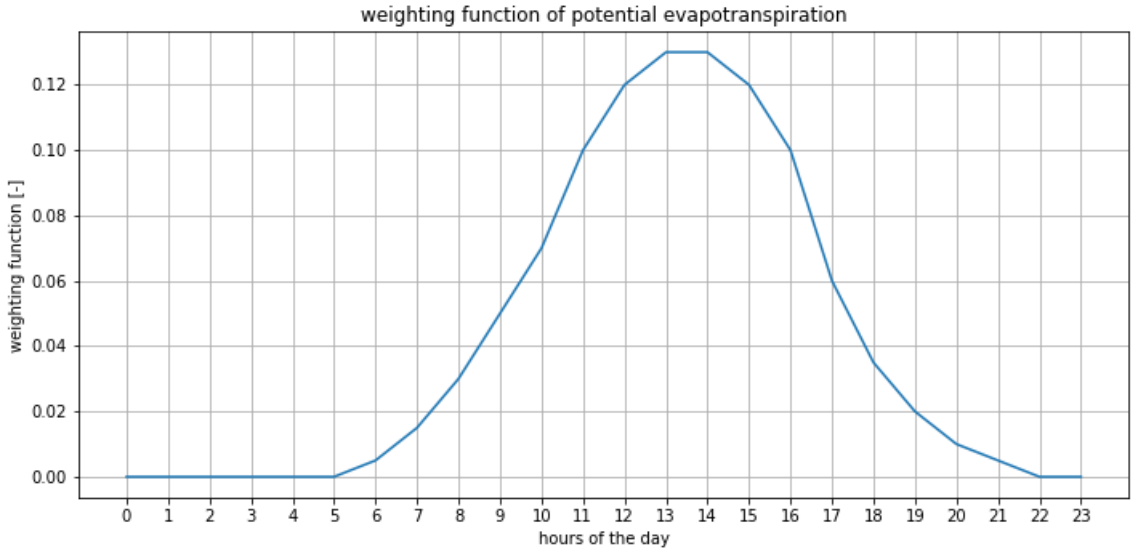


Figure 22: Weighting function used for calculating the hourly potential evapotranspiration

PET calculated with Hargreaves, MOD16 PET and AET data were compared. This was done by plotting a timeseries of the evapotranspiration products. Additionally, the annual actual evapotranspiration amount from MOD16 and modelled by HYPE was evaluated.

4.2.5 Calibration

The next step for setting up HYPE was to develop and choose an automatic calibration method, that is able to create reproduceable results. Two calibration methods provided by HYPE were tested and compared. In the next step, soil thickness and catchment characteristics were varied and calibrated. Another approach was evaluated, where the calibration was divided into different processes. For calibrating catchments with more than one gauge, a method was tested and chosen and as a last step, calibration methods for improving low flow conditions were tested.

4.2.5.1 Calibration methods

HYPE provides nine different automatic calibration methods. They are based on Monte Carlo, Differential Evolution Markov Chain, Brent, the method of steepest decent and quasi-Newton (SMHI 2021). In this thesis, two of these methods, DEMC and a progressive Monte Carlo simulation were tested, because these two methods showed a high efficiency with good model performances in Schönfelder and Baclet (2022), where all of these calibration methods were evaluated. The KGE was applied as objective function for the evaluation.

To compare the calibration methods, the number of simulations was varied and the performance of each run compared. A calibration was then repeated to test whether the same performance was achieved again. To also test the efficiency of the methods, the wall time (time, that the calibration takes) was measured and compared. All of these tests were only done in one catchment (Nausta) and with one dataset (seNorge2), assuming that they will not vary for other areas. Different catchment calibrations were also tested by Saldaña Espinoza (2022). Each catchment showed similar behaviour towards the calibration method. More simulations resulted in better KGEs as well as a more similar performance when repeating the calibration.

Regarding DEMC, an error was found in the HYPE-forum (Pers 2022a) with the explanation that the optimal dataset of this calibration is not saved in `respar.txt` (file that saves the resulting optimal parameter values (SMHI 2021)), but instead has to be extracted from `bestsims.txt` (saves the parameters of a number of best simulations (SMHI 2021)). Since this was found out only late in the study, this knowledge could not be included into decision making for the calibration strategy. After the discovery, further tests with DEMC were done anyways.

4.2.5.2 Stepwise calibration

Additionally to a one-time calibration, a stepwise calibration was tested. This method is also described in Arheimer and Lindström (2013) where HYPE was calibrated for Sweden. The main idea is to take one process at a time and calibrate the dependent parameters in each step. In this study, they tried to follow the water flow path from upstream to downstream. A stepwise approach was also done when setting up the World-Wide-HYPE model (Arheimer et al. 2020). Interesting here was that processes were partly just calibrated in areas where they are relevant. They also divided the catchments into different climate zones. There, eleven processes were defined and calibrated iteratively. Saldaña Espinoza (2022) used a stepwise approach for calibrating different catchments and improving lake outflow calibration. The results showed an improved KGE when doing a stepwise calibration. In this approach, the water level of lakes was calibrated in one of the steps additionally. In this study, the step wise calibration was tested on one of the catchments at first and showed promising results, because of an increase of the model performance after each step and a better resulting KGE than with any other calibration method.

To avoid overcompensating errors with different processes, the initial calibration (with all parameters) was set up to already have a high number of simulations, to provide stable results for each parameter set. That way, the stepwise calibration should just lead to an improvement of

parameter sets (fine tuning) instead of changing the model dynamic of the whole catchment. In the final version, the following processes and their parameters were calibrated:

- Soil: *wcfc1, wcfc2, wcfc3, wcwp1, wcwp2, wcwp3, wcep1, wcep2, wcep3, mperc1, mperc2, sfrost, rrcs1, rrcs2, rrcs3, srrate, macrate, mactrinf, mactrsm, frost, srrcs, surfmem, ttrig, treda, tredb, depthrel*
- Winter: *ttpd, ttpi, deepmen, cmlt, ttmp, fscmax, fsck1, fsceff, fscdistmax, fscdist0, fscdist1, sdnsnew, snowdensdt, fsclim, fsckexp (pcusnow)*
- Divers: *lp, epotdist, cevp, rivvel, damp,deadl, deadm, (pcurain)*
- Ilakes: *icatch*
- Olakes: *gratk, gratp, grata*

For an explanation of parameters see Appendix A.2. *pcurain* and *pcusnow* were only calibrated when precipitation correction was applied. The stepwise approach was done iteratively, meaning that the calibration of the processes was repeated another time. When the KGE decreases during one step of the calibration, the resulting parameters are not applied into the next step.
The ilake calibration was done in another way than the rest of the processes. Since the parameter *icatch* can be defined global (same value for each sub-catchment) or regional (value defined in `GeoData.txt` and different in each sub-catchment), *icatch* was first calibrated for all catchments (in the initial calibration) and afterwards changed directly in the `GeoData.txt` file. For that, a Python script was developed that used a Monte Carlo approach: First the text file is read, the column "*icatch*" for each sub-catchment is changed with a random number between defined boundaries and then HYPE is executed. The results are then read and saved and then another set of parameters is calculated randomly again. These boundaries were set for each catchment manually.
One exception was Gaula with 250 sub-catchments. Because a manually set of boundaries would be too much effort, an approach suggested by Schönfelder (2017) was applied. Here, the size of the lake (A_{lake}), the distance between the lake and the sub-catchment outlet (d_{outlet}) and to the center (d_{center}) are set into relation with the sub-catchment area (A_{basin}). Then, two parameters (m,u) are varied between 0 and 1 and *icatch* calculated. This *icatch* is than calibrated by varying m and u.

$$icatch = \left(u \times \frac{\sqrt{A_{basin}}}{d_{outlet}} + m \times \frac{\sqrt{A_{basin}}}{d_{center}} \right) \times \frac{A_{lake}}{A_{basin}} \qquad \textit{Eq. 33}$$

For sub-basins with more than one lake, the one with the highest drainage area was chosen. Lakes smaller than 0.05 km^2 were not considered. For these lakes, a random value between 0 and 0.05 was used as *icatch*, *olakes* have a value of 1.

The following algorithm was used for the stepwise calibration:

1) Initial calibration with all parameters
   - Calibration method: Progressive Monte Carlo
   - num_bpmax: 5
   - num_bp: 5000
   - num_ens: 5
2) Calibrate soil processes
   - Calibration method: Progressive Monte Carlo
   - num_bpmax: 10
   - num_bp: 200
   - num_ens: 5
3) Calibrate snow parameters
   - Calibration method: Progressive Monte Carlo
   - num_bpmax: 10
   - num_bp: 200
   - num_ens: 5
4) Calibrate all other parameters
   - Calibration method: Progressive Monte Carlo
   - num_bpmax: 10
   - num_bp: 200
   - num_ens: 5
5) Calibrate ilakes
   - Calibration method: Monte Carlo
   - n: 500
6) Calibrate olakes
   - Calibration method: Progressive Monte Carlo
   - num_bpmax: 10
   - num_bp: 200
   - num_ens: 5
7) Repeat steps 2-6

4.2.5.3 Handling catchments with multiple discharge gauges

As seen in Figure 12, there are seven discharge stations available in Gaula, that have available time series in the time period of this study. Two of them are only measured in a daily timestep (Gaula ovf. Fora and Killingdal). To develop a calibration strategy, different variations of station-calibrations were tested.

First of all, the calibration was done using the median KGE of all gauges (median all stations). For that every station was calibrated and the objective function was to maximizes the median KGE of all stations. The same procedure was repeated with the average KGE (average all stations). After that, specific stations were calibrated together using the average KGE as objective function, and the parameter sets were then applied to the whole catchment. Finally, every station was calibrated by itself. When calibrating the whole catchment, the number of simulations was 2000, when the gauge had a small upper catchment, the calibration was done with a higher number of simulations (25,000 simulations, according to the initial calibration of the stepwise approach). This comparison was done using seNorge2. The results were than compared and analysed to decide which strategy will be used for the final calibration of Gaula.

4.2.5.4 Low flow calibration

Although the right forecast of flood discharges is useful in a model (for example for flood forecasts or risk assessment), low flows are also necessary to regard. For example, fish populations depend on a minimum water level to breed and live in a river. Artificial decrease of water levels in rivers may have strong impacts on local fish population. It is known, that low flows tend to be underestimated in the Norwegian HYPE model (Schönfelder (2019), Schönfelder and Baclet (2022)), different methods to improve their performance were tested. The question of interest was whether precipitation products are responsible for this behaviour.

Different approaches to improve the model performance were applied. First discharge data was extracted to only include low flows. They were defined as the 20 % lowest flows during the measurement period. These values were then applied for the calibration. Another approach was using different objective functions. Here, the relative bias and root mean squared error were tested. To decrease the number of calibrated values, parameters with higher sensitivities towards low flows were detected by changing parameters one at a time and looking at the change of the KGE. For these parameters, new boundaries were defined, that created better low flow performances. This evaluation was done for one catchment using both datasets. The results of the analysis were then transferred to another catchment to validate the results.

4.2.5.5 Soil layer and soil thickness calibration

Since soil layer thickness cannot be included in an automatic calibration, a pre-analysis of different soil layer thicknesses was done. For that, the variables *stream-depth*, *soil-layers* and *depth 1-3* in `GeoClass.txt` were changed manually, and the influence of layers was analysed. The progressive Monte-Carlo method with 200 simulations per stage and 10 stages was used for that. The reason for that is the lower computing effort when using a smaller number of simulations, but still a qualitative result could be created. After calibrating different soil layer thicknesses, the results with the best KGE were considered to be the right soil thickness. The stream-depth is another parameter that can be changed manually. This parameter describes the height of the ground water participating in the flow and standing water underneath. To test this dependence, the best results of the soil layer calibrations were afterwards calibrated with varying stream depths. The results of these calibrations can be found in chapter 5.2.3.

4.2.6 Summary of modelling process

This section is a summary of the steps that were done to create and calibrate the model of each catchment. The used scripts are included in the digital appendix. A description of the programmes used is in Appendix I. The following steps were done for each catchment:

1) Delineation:
   All of the catchments already had a delineation with sub-basins. Still, they had to be altered, when a gauge station was not at the outlet, an olake was included or to increase the degree of detail. These small adjustments were made with WHIST.
2) Generate input data:
   Gridded precipitation and temperature data is saved in netcdf4 files. These files contain matrixes of meteorological values with dimensions of time, latitude and longitude. Using a geodata file to extract data with a R script. The netcdf4 files are saved on a download server and were extracted online without a previous download.
3) Generate model set up files:
   For a functioning HYPE model, a directory with specific files has to be created. The HYPE model set up is explained in Appendix A.1. Most of the files are similar in each catchment or have to be readjusted in a small manner. `GeoData.txt` had to be recalculated when the delineation was changed using QGIS (slope, mean elevation and SLCs).
4) Soil layer calibration:
   Before starting the actual calibration, each catchment was modelled with a different number of soil layers and different thicknesses.
5) Calibration:
   After finishing step 1-4, the calibration process was started. Each catchment was calibrated according to the stepwise calibration described in section 4.5.2. This method combined the automatic calibration from HYPE with a Python script that processed the stepwise calibration. Each catchment was calibrated at least four times: two times with the seNorge2 dataset (with and without precipitation correction) and two times with NWP (with and without precipitation correction). In two catchments (Usma and Grunnåi) two further calibrations were done with the weather station data as precipitation input. Because of the size of the Gaula catchment, the initial calibration of Gaula was decreased to 10,000 simulations and the stepwise calibration was not iterated a second time.
6) Validation
   After calibration, the resulting parameter sets were validated for a different time period (see Figure 19) to check whether the results are reliable. Because discharge data for Usma was not available for a long time period, it was not validated.

4.3 Result analysis

After setting up and calibrating the models, the results were compared and analysed. At first, model performances of each catchment and different precipitation methods from calibration and validation were compared to find out which product created the best model performance. Then, performances of the model in summer and winter, low flow behaviour and the correlation between simulated and observed discharge were compared. The model results of each catchment were then evaluated by visualizing the discharge range using different products and cumulative volume sums. In a next step, water balances of each catchment were created, comparing the precipitation volume with discharge and evapotranspiration. In that process, the evapotranspiration amount modelled by HYPE was compared with MOD16 data. Additionally, seasonal behaviour of the water balance was analysed. To detect whether HYPE creates unnatural water accumulations or losses, the water

storages were checked as well. For that, the model setting *printwaterbal* was applied. With that, the model creates additional output files containing the water storage amount of each catchment and for each timestep. This information was used to analyse the water storage changes and visualize possible storage accumulations or losses.
At last, specific events in each catchment were chosen and analysed according to precipitation data and modelling results. For that, events were identified that showed mismatching behaviour of the model compared to the discharge data. These events were then visualized spatially over time and the simulated discharge analysed. For these events, an inverse modelling approach (Kirchner 2009) was applied to test whether they are comparable to the rain rate suggested by the datasets.

4.4 Sensitivity analysis

In the last step of this study, the model accuracy and correctness was tested by doing a sensitivity analysis. Three different approaches were used for that. First, GLUE was applied for the calibration results, the calibration and validation period were switched and in the last step, the physicality of parameters was analysed and evaluated.
GLUE (see chapter 2.1.2) was applied for two catchments. To create different calibration results and parameter spaces, the original approach was adapted, because the calibration method was decided to be progressive Monte-Carlo instead of a general Monte-Carlo. To get a sufficiently large enough number of different best-parameter results, the number of ensembles could have been increased, but that would have led to a change in the calibration method. That is why it was decided to create different results by repeating the calibration a certain amount of times. In (Sivasubramaniam et al. 2020), the parameter space of the best 50 results were analysed, here only the best 20. This led to a repetition of the calibration four times with five outputs per run. Instead of the stepwise calibration, only the initial calibration was used for this comparison. GLUE was applied for only two catchments, Grunnåi and Nausta. First, the parameter space was visualized by normalizing (n_i) the values (p_i) according to their boundary values (p_{min} and p_{max}) of the optimization.

$$n_i = \frac{p_i - p_{min}}{p_{max} - p_{min}} \qquad \textit{Eq. 34}$$

This can give information about whether the chosen boundaries are suitable or whether the parameters converge against one of the boundaries. Each calibration result was further visualized individually. Here, the best five results of each of the four calibrations were divided by colours. That way, differences between each calibration could be found more easily.
Another step to test the accuracy and suitability of the model was to switch the calibration and validation period and evaluate the performance change. This means that for seNorge2, that is calibrated between 2013/09 and 2016/09 (time period 1) and validated between 2011/09 and 2013/09 (time period 2), the times were switched (calibration in time period 2 and vice versa). The same was done with NWP, which is originally calibrated between 2014/09 and 2017/09 (time period 1) and validated between 2017/09 and 2019/09 (time period 2).
The last step was to compare physical parameters. For that, *pcurain* and *pcusnow* were compared in each catchment and with each dataset. Additionally, the river velocity (*rivvel*) was compared in value and was also set into comparison with the slope and elevation of a catchment. To see whether the SLCs have a realistic behaviour, land use dependent parameters *cmelt* and *cevp* were analysed as well.

5 Results

In this chapter, the results of this study will be provided. At first, the precipitation products were generally compared. Afterwards, different calibration methods were evaluated. After running the model, the results were analysed and compared, as well as water balances and storage changes. Additionally, selected events were analysed. At last, a sensitivity analysis of the model was done using GLUE, switching the calibration period and evaluating specific parameters.

5.1 Comparison of the datasets

In the following chapter, the results of comparing the two datasets are described. To get a general overview over the spatial distribution of precipitation accumulated annual precipitation sums of the datasets were calculated. To compare the temperature datasets, the measured temperature of each catchment was compared by using a linear regression and correlation. Temperature and precipitation were compared with a weather station owned by a hydro power company as the ground truth. At last, the precipitation-runoff relation of each catchment was compared.

5.1.1 Spatial variability of precipitation

In Figure 23 the annual precipitation sums of 2014 and 2015 over the Gaula catchment are visualized. Notice that the year was defined from September to September. The same figures of other catchments can be found in appendix C.1.

NWP precipitation sums of Gaula varie between 788-1658 mm/a with higher precipitation in the higher elevated areas. SeNorge2 varies between 477-1208 mm/a with the highest precipitation in the lower part of the basin. Radar data assumes the annual precipitation between 225-904 mm/a with higher precipitation in the lower basin as well. In Usma, the annual precipitation according to NWP varies between 1,093-1,421 mm/a with higher precipitation in the west side of the basin. SeNorge2 is more homogeneous and varies between 702-772 mm/a. Radar data is lower than the other products with 440-619 mm/a. It is also mostly homogeneous over the catchment. The NWP product over Nausta varies between 2,263-4,205 mm/a. The highest precipitation is located at the higher elevated sub-catchments. SeNorge2 however has the highest precipitation in the lower elevated ones, varying between 2,066-3,723 mm/a. Radar data has a beam blockage in the upper catchment. It varies between 180-795 mm/a. Precipitation in Grunnåi varies between 1,335-2,110 mm/a (NWP) and 1,167-2,016 mm/a (SeNorge2). Both datasets have the highest precipitation in the upper catchment, located in the north of the basin. Radar data is again lower between 588-807 mm/a. The highest precipitation is located in the south-east of the catchment. In Surna, precipitation varies between 1,170-2,800 mm/a (NWP) or 750-1,500 mm/a (seNorge2). The radar measures precipitation of about 422-952 mm/a. Both datasets have the highest precipitation in the south-west of the catchment.

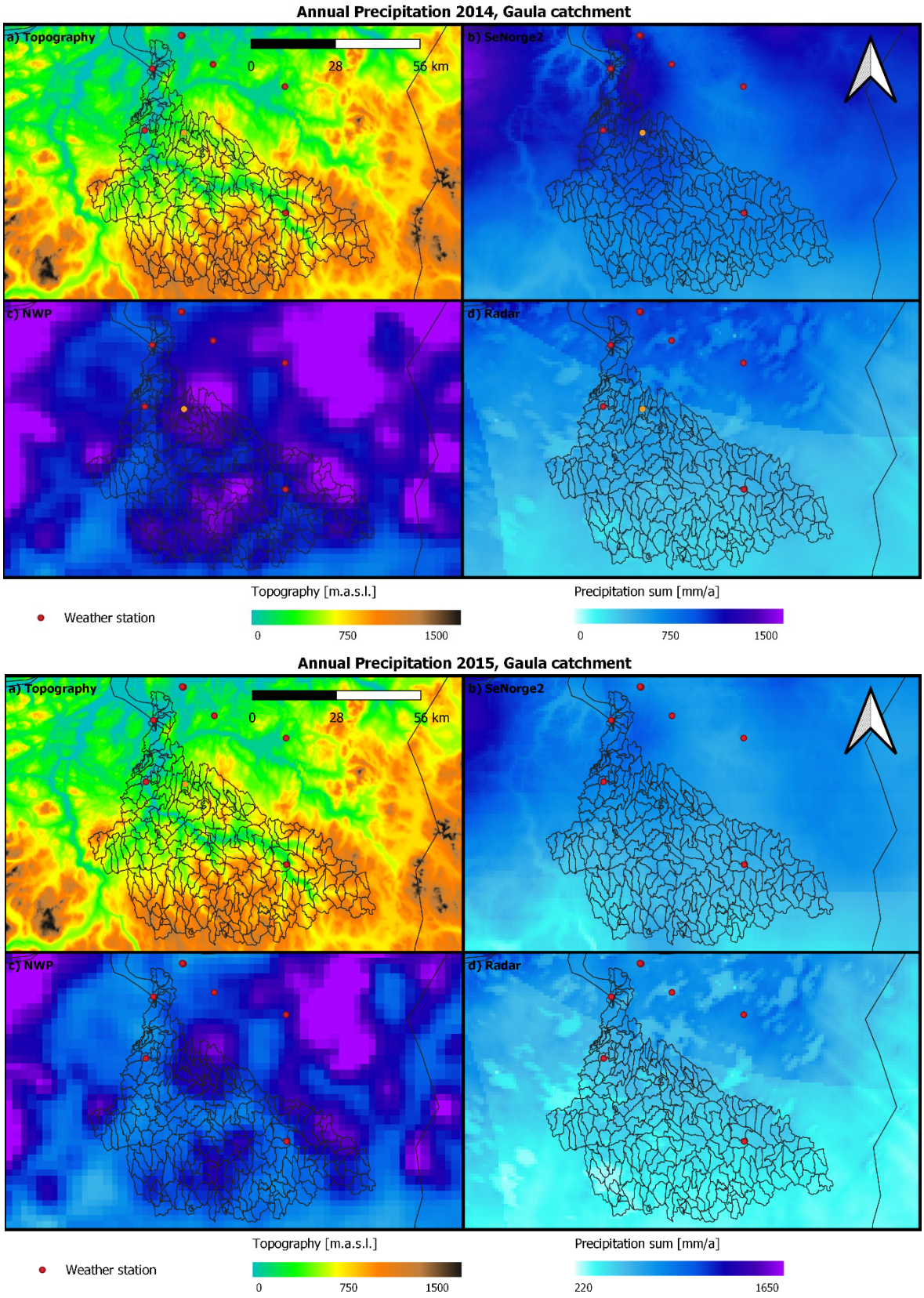


Figure 23: Annual Precipitation sum of September 2014 to September 2015 (upper figure) and September 2015 to September 2016 (lower figure) over Gaula

An annual precipitation sum was further calculated for the KliNoGrid dataset (combination of radar and gauge data). Figure 24 shows the precipitation sum of KliNoGrid over Norway and Gaula.

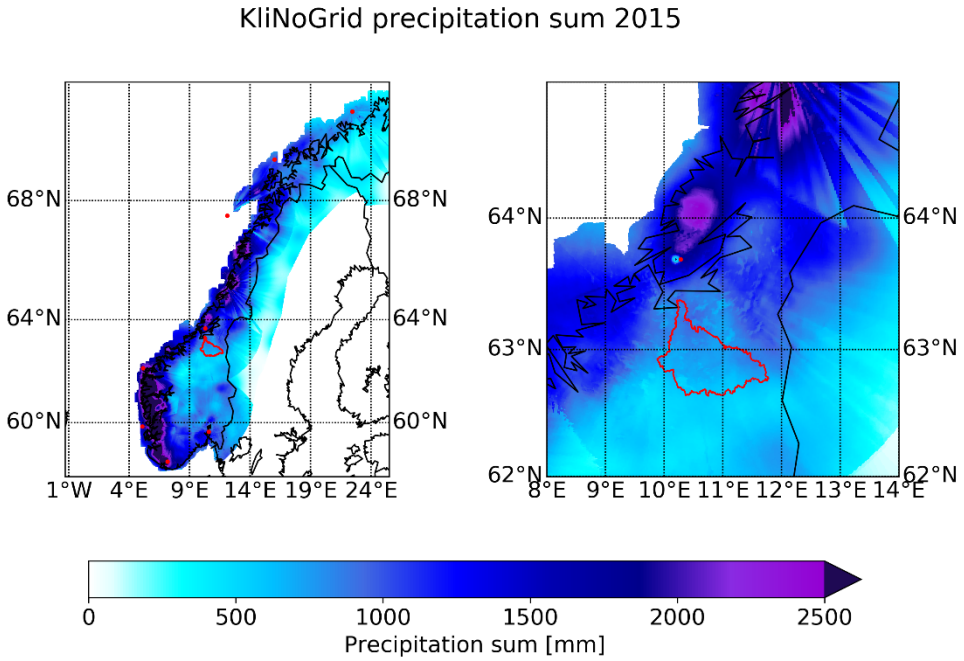


Figure 24: KliNoGrid precipitation sum over Norway (left) and Gaula (right) of 2015. The red dots are locations of radar stations.

5.1.2 Spearman correlation

As a next step, the Spearman correlation was calculated. Lumped temperature and precipitation was calculated with seNorge2 and NWP. Afterwards, the Spearman correlation of the overlapping time period (09/2013 to 2017) was calculated. The results can be found in Table 6. To consider the spatial variability of the datasets, the Spearman correlation was further calculated for each sub-catchment of the basins. These results were then compared with the mean elevation of the sub-catchments. First, all precipitation types were considered at once (see Figure 62, Appendix C.2). To get more information about the influence of snow, mixed precipitation and rain, the observations were divided by temperature. The comparison is visualized in Figure 25.

Table 6: Spearman correlation of lumped catchment temperature and hourly and daily precipitation

	Spearman Correlation of Temperature	Spearman Correlation of Precipitation (hourly)	Spearman Correlation of Precipitation (daily)
Gaula	0.97	0.67	0.89
Nausta	0.95	0.61	0.81
Usma	0.94	0.54	0.80
Surna	0.99	0.66	0.86
Grunnåi	0.97	0.48	0.76

In general, temperature data of seNorge2 and NWP has a high correlation (all over 0.94). Hourly precipitation has a lower correlation than when aggregated to a daily timestep. The smallest catchments (Usma and Grunnåi) have the lowest correlation. When looking at the sub-catchment scale, the correlation is slightly decreasing with higher elevations. The correlation between hourly precipitation is in general relatively small, with a better correlation when aggregating them up to daily precipitation. The correlation of precipitation is decreasing with the elevation as well. Cold

temperatures lead to a decrease of correlation, especially in Nausta and Surna. Snow (temperatures below -5 °C) change the influence of elevation and catchment. Here, the correlation is constant over elevation, in some cases (for example Surna) higher elevations have better precipitation correlations.

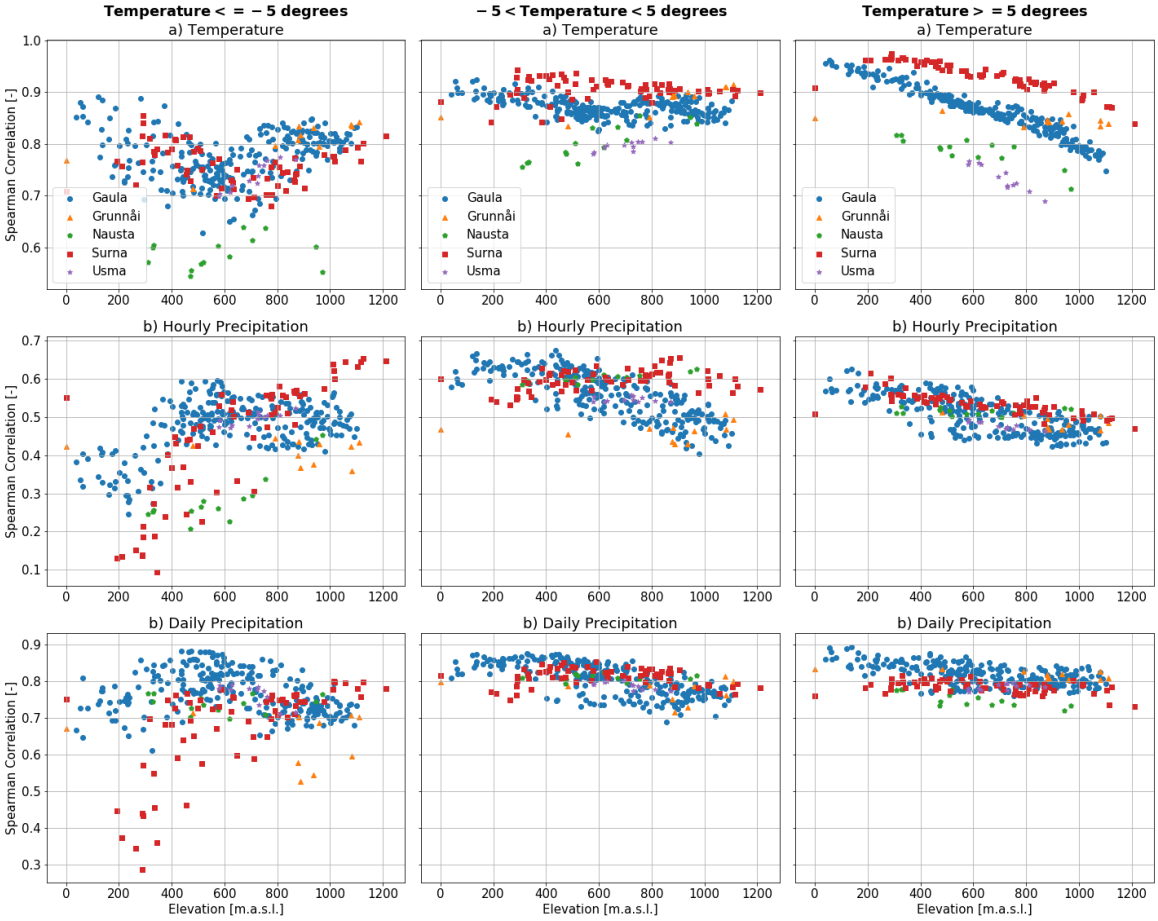


Figure 25: Spearman correlation of seNorge2 and NWP of each sub-catchment sorted by mean elevation and temperature

5.1.3 Weather station

A comparison of the datasets is complicated because there is not a lot of information about the ground truth available in most of the catchments that is not already included in the products itself. Nevertheless, a weather station maintained by TrønderEnergi AS is available that is located 30 km away from the Usma catchment. That is why the two datasets were compared with that weather station. To compare the observations, the gridded datasets were extracted at the location of the weather station and timeseries created. The comparison was done between 2014 and 2017.

At first, temperature and precipitation were compared by using the correlation and a linear regression (see Figure 26). Temperature has a higher correlation than precipitation. When looking at the linear regression, the temperature seems to be slightly underestimated by seNorge2, especially lower temperatures. NWP only underestimates temperatures above 0 °C. Hourly precipitation has a low correlation. Both datasets seem to underestimate the precipitation amount. When aggregating precipitation to daily values, the correlation increases strongly. NWP seems to overestimate the precipitation, whereas seNorge2 underestimates it.

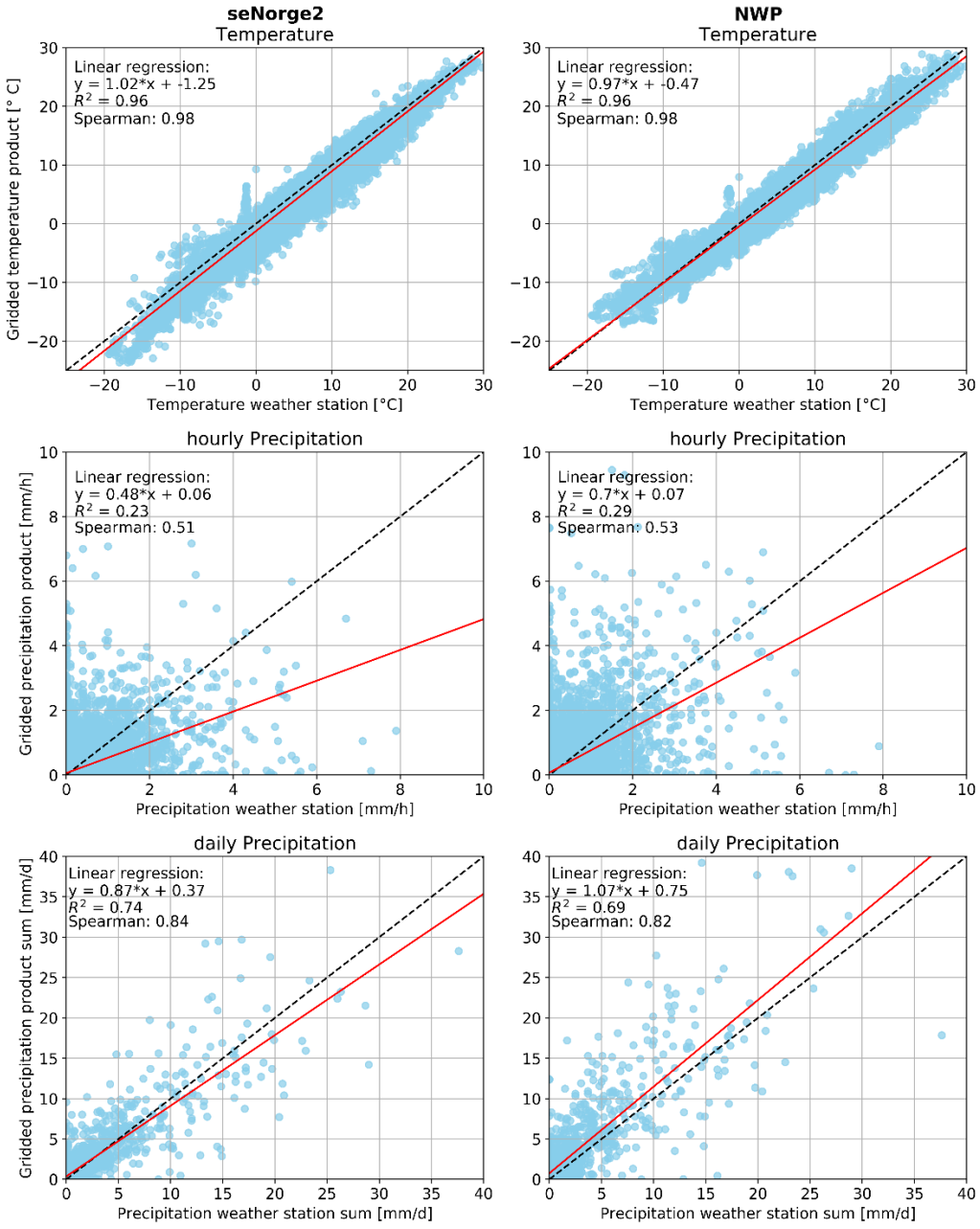


Figure 26: Comparison of NWP, seNorge2 and a weather station near Usma. Time period: 2014-2017

In a further step, the mean temperature and precipitation was calculated of the weather station, NWP and seNorge2. This provides information about biases of the datasets. The results are summarized in Table 7. The mean annual temperature, as well as in summer and winter, is assumed lower in the gridded datasets. NWP is slightly closer to the gauge observation. The precipitation sum is estimated higher in both datasets than measured by the gauge. NWP is 38 % higher than the observed precipitation amount (31 % in summer and 28 % in winter). SeNorge2 overestimates the precipitation by only 2 % with almost no difference in summer and 7 % in winter.

Table 7: Mean Temperature and Precipitation of weather station, NWP and seNorge2 near Usma

Temperature			
	Mean annual temperature [°C]	**Mean temperature in summer* [°C]**	**Mean temperature in winter** [°C]**
Weather station	4.87	9.50	0.23
seNorge2	3.72	8.57	-1.15
NWP	4.25	8.99	-1.15
Precipitation			
	Mean annual precipitation sum [mm]	**Mean precipitation sum in summer [mm]**	**Mean precipitation sum in winter [mm]**
Weather station	906	502	404
seNorge2	926	503	423
NWP	1246	730	516

*summer between April and September
**winter between October and March

5.1.4 Mass balance

A way to evaluate the quantity of precipitation is to compare the volume of discharge, precipitation and evapotranspiration. For that, the discharge volume of observations was calculated within a hydrological year (September to August) and was scattered against the precipitation volume of the basin. Figure 27 combines the water balance of all catchments. Dots over the black line can be assumed to underestimate precipitation, points underneath it are overestimating. The difference between points and the black line are summarized as boxplots (including each year) at the lower right side.

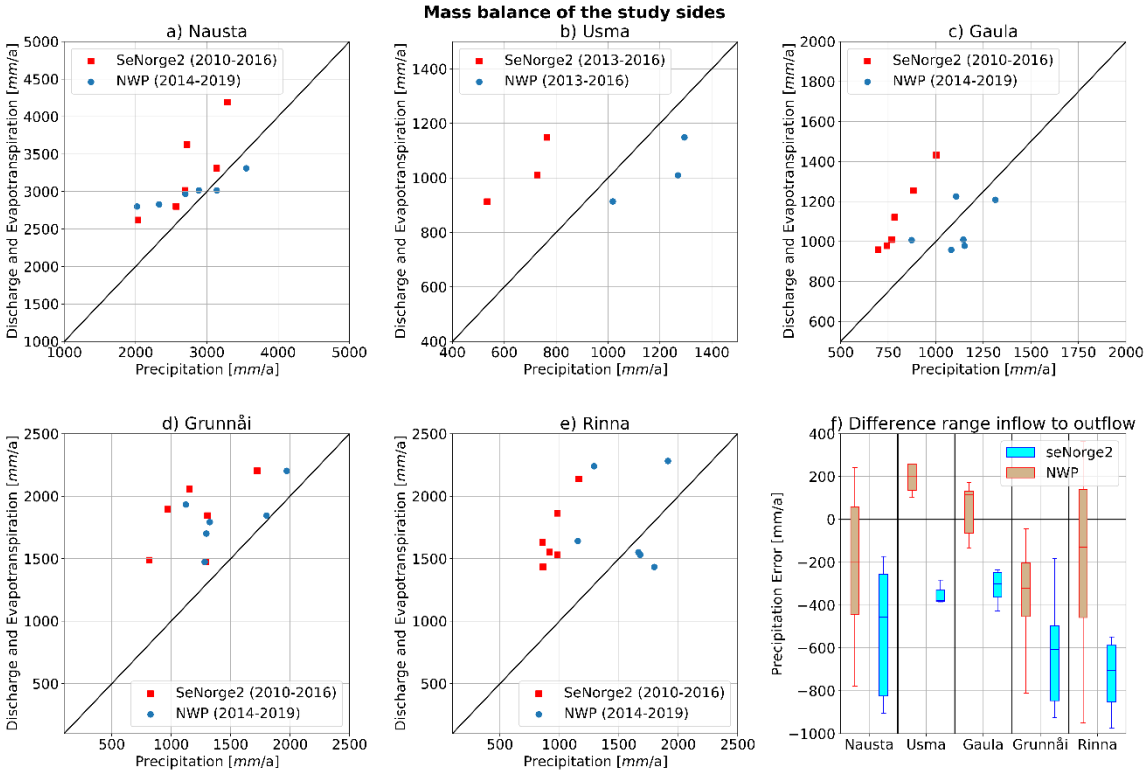


Figure 27: Comparison of annual volume amount of runoff plus evapotranspiration (MOD16) and precipitation for each catchment and differences to the water balance (lower right)

5.1.5 Inverse modelling

The last step of comparing the dataset was to use an inverse hydrological model as a comparison. For that, the precipitation was calculated from discharge data according to chapter 2.1.3. The calculation of sensitivity-curve-parameters is shown in appendix D. Because snow melt in the basin could bias the results, a time period between June and October was modelled between 09/2014 and 09/2016. The results of these two years were then averaged for each year. This was done for each catchment. The resulting precipitation amount of the inverse model are compared with the datasets in Table 8.

Table 8: Precipitation amount between June and October estimated by an inverse hydrological model (Kirchner 2009), seNorge2 and NWP. Data is averaged between 2014 and 2015.

	Kirchner (2009)	seNorge2	NWP
Nausta	1066 mm	603 mm	729 mm
Usma	934 mm	306 mm	447 mm
Gaulfoss	437 mm	305 mm	422 mm
Grunnåi	1289 mm	728 mm	620 mm
Rinna	718 mm	393 mm	616 mm

5.2 Model calibration

This chapter combines the results of testing different calibration methods as well as the results from the actual calibration of each catchment and each precipitation product.

5.2.1 Development of a calibration strategy

To create reliable results that can provide information about the datasets, the model must create reasonable outputs. A good calibration is the key for that. In this study, different approaches of calibration were tested and evaluated. First, the automatic calibration method was evaluated, soil layer calibrated for each catchment and a stepwise calibration applied. For further analysis, low flows were calibrated, different snow models tested and a method for calibrating multi-gauge catchments developed.

5.2.1.1 Automatic calibration method

In this section, two automatic calibration methods were tested and compared. They were both applied in Nausta with the seNorge2 dataset. Different combinations of parameters (generations and populations) were tested. The calibration period is from 09/2013 to 09/2015. Validation is between 09/2015 and 09/2016.

DEMC was tested first. For that, the number of generations and populations was varied for the calibration. The calibration was repeated two times to test the reproducibility. It needs to be mentioned that these results were only archived at the end of this study. When first testing this method, the validation results were lower (the highest value had a KGE of 0.79). After gaining more information about the output variable structure (explained in chapter 4.2.5.1), the results according to Table 9 were archived. The best KGE is achieved with the highest number of simulations. The validation shows good agreement as well.

As a quick test to see whether DEMC would have outperformed the chosen calibration method, it was applied with 9000 simulations in two catchments (Nausta and Usma) to compare the

performance. In Nausta, DEMC created a KGE of 0.923 with seNorge2 in the calibration period compared to 0.927 that was calibrated with a stepwise approach. NWP performed with a KGE of 0.850, the chosen method (stepwise) had an KGE of 0.870. For Usma and seNorge2 DEMC calibrated a KGE of 0.757 (stepwise 0.797) and with NWP 0.866 (stepwise 0.876).

Table 9: Summary of different calibration tests using DEMC

Calibration parameters	KGE calibration	KGE validation	KGE calibration	KGE validation	Wall time*
	Calibration 1		Calibration 2		
ngen: 20 npop: 100	0.820	0.817	0.816	0.799	2.5 h 11 min/sub-basin
ngen: 100 npop: 20	0.894	0.857	0.852	0.897	2 h 9 min/sub-basin
ngen: 200 npop: 20	0.887	0.855	0.908	0.892	3.5 h 15 min/sub-basin
ngen: 300 npop: 30	0.918	0.895	0.924	0.892	20 h 86 min/sub-basin

*CPU: Intel® Core™ i7-9850H CPU @ 2.60 GHz

Figure 28 shows the development of KGE during the calibration process with 2000 and 9000 simulations. At the beginning, the KGE varies between zero and one, but after enough simulations, the KGE starts converging against a higher value. However, even after 9000 simulations the KGE didn't fully converge.

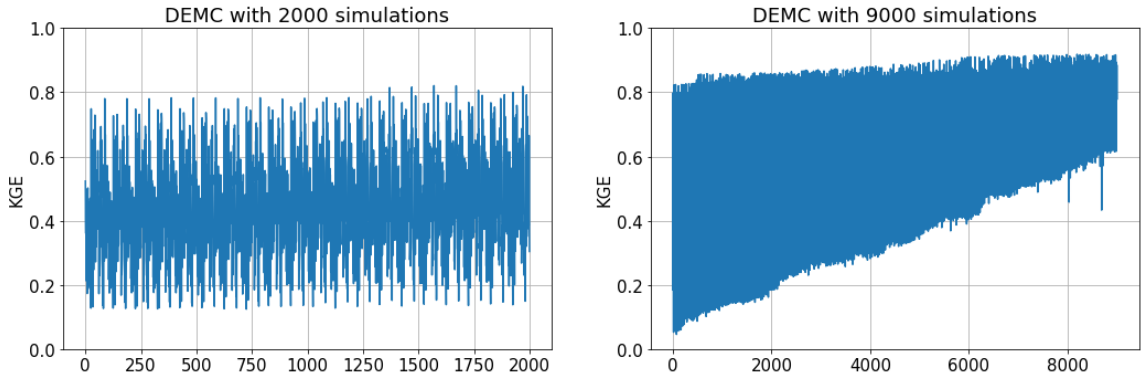


Figure 28: KGE development during the DEMC calibration with ngen=100 and npop=20 (left) and with ngen=300 and npop=30 (right)

In a next step, the progressive Monte Carlo Method was applied for calibration. Here, the number of stages and simulations per stage were varied. Results are summarized in Table 10. Because the progressive Monte Carlo method depends on results of the first stage, the calibration was repeated a second time, to compare performances of each run and test the stability of the results. It can be seen that the highest KGE in calibration and validation was archived with the highest number of simulations. The change between the first and second calibration is the lowest in that case.

Table 10: Summary of different calibration methods using the progressive Monte Carlo method

Calibration parameters	KGE calibration	KGE validation	KGE calibration	KGE validation	Wall time*
	Calibration 1		Calibration 2		
num_bpmc: 200 num_bpmax: 5 num_ens: 5	0.823	0.785	0.838	0.824	1h 4 min/sub-basin
num_bpmc: 200 num_bpmax: 10 num_ens: 5	0.871	0.855	0.853	0.821	1.7 h 8 min/ sub-basin
num_bpmc: 200 num_bpmax: 15 num_ens: 5	0.833	0.816	0.865	0.852	2.5 h 11 min/ sub-basin
num_bpmc: 1000 num_bpmax: 10 num_ens: 10	0.837	0.841	0.864	0.837	17 h 73 min/ sub-basin
num_bpmc: 5000 num_bpmax: 2 num_ens: 2	0.864	0.848	0.852	0.814	19.5 h 84 min/ sub-basin
num_bpmc: 5000 num_bpmax: 5 num_ens: 5	0.869	0.866	0.869	0.864	25 h 107 min/ sub-basin

*CPU: Intel® Core™ i7-9850H CPU @ 2.60 GHz

5.2.1.2 Stepwise calibration

A stepwise calibration was first applied in the Nausta catchment and showed an improvement of the KGE, when done as explained in chapter 4.2.5.2. In the first test case, the initial calibration was done with only 2,000 simulations. The improvement of the KGE during the stepwise calibration was from 0.69 to 0.80 and from 0.91 to 0.94 for an already high KGE. This fine-tuning of parameters was decided to be applied for all sub-catchments.

An example of a stepwise calibration process is shown in Figure 29 and Figure 30. When calibrating a model with NWP, the KGE is improving slightly and the KGE of the validation is improving as well. In the second iteration, there is no or only a small amount of improvement. SeNorge2 shows a higher increase during the stepwise calibration. Nevertheless, the KGE of the validation is varying within each step and only slowly increasing.

The results of other catchments are summarized in Appendix D and figures of stepwise calibrations are in the digital appendix. In general, the KGE is improved in the final calibration between 0.01 and 0.07. The KGE was more or less constant after the first iteration. The performance of the validation was only decreased in one catchment (namely Surna) but was constant or increasing in every other catchment.

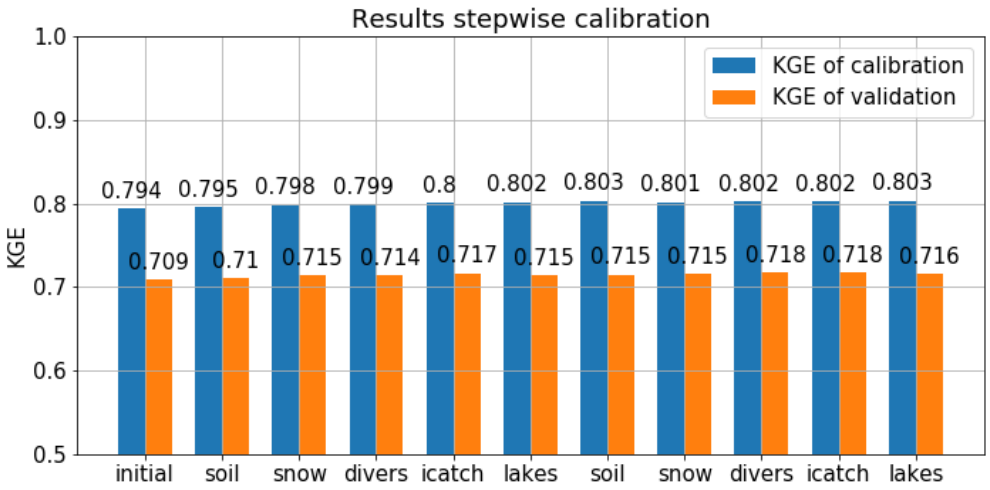


Figure 29: Stepwise calibration results in Grunnåi with NWP

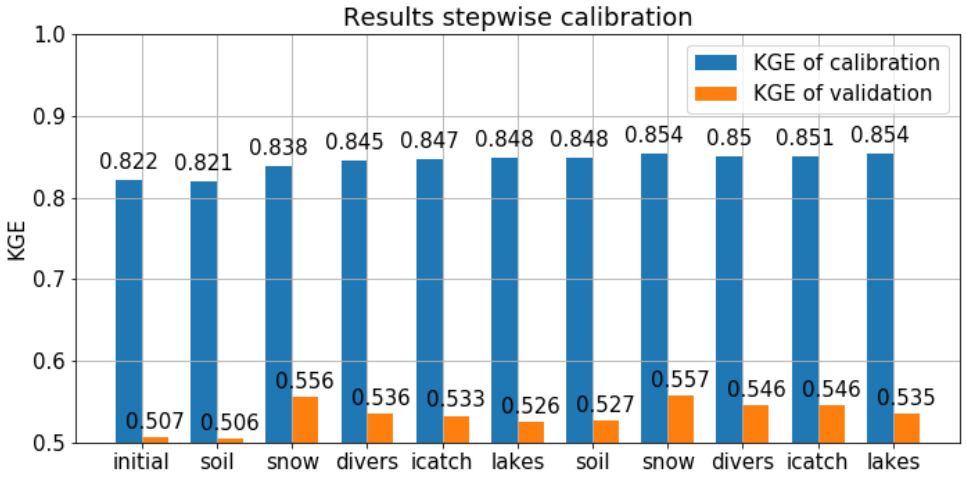


Figure 30: Stepwise calibration results in Grunnåi with seNorge2

5.2.1.3 Multi-gauge calibration

Another strategy had to be developed for Gaula, because it contains more than one discharge gauge that can be used for calibration. Table 11 summarizes the results of this process. Here, different stations were calibrated (all of them or only three/two/one). The KGE of the calibrated stations is given, as well as the KGE of all stations when using the calibrated parameter set. The time and number of calibrations is included as well. In the first row, the KGE is given as the mean of all stations, in the other rows it is the average.

Table 11: Results of different calibration strategies in Gaula

	KGE calibration	KGE all stations	KGE Validation	KGE Validation all stations	Time [h]	Number of calibrations
Median all stations	0.632	-	0.642	-	25 *	
Average all Stations	0.679	-	0.483	-	25 *	2000
Gaulfoss-Lillebudal-Gaua	0.703	0.585	0.464	0.464	47 **	2000
Gaulfoss-Hugdal	0.635	0.635	0.581	0.581	48 **	2000
Gaua	0.842	0.704	0.73	0.649	8 *	25000
Lillebudal	0.667	0.664	0.487	0.575	14 *	25000
Killingdal	0.686	0.638	0.642	0.553	15 *	25000
Hugdal bru	0.737	0.682	0.733	0.656	135 **	25000
Eggafoss	0.755	0.636	0.703	0.547	48 *	25000
Gaula ovf. Fora	0.78	0.606	0.651	0.505	275 **	25000
Gaulfoss	0.750	0.617	0.7496	0.544	16 *	2000

* CPU: Intel® Core™ i7-9850H CPU @ 2.60 GHz

** CPU: Intel® Core™ i7-8665U CPU @ 1.90 GHz

After calibrating the catchment with these methods, the KGE of all the other stations was checked. Figure 31 shows the KGE distribution of all gauges when calibrating with the median and average KGE. It also shows the distribution when only one catchment is calibrated (Gaua and Gaulfoss).

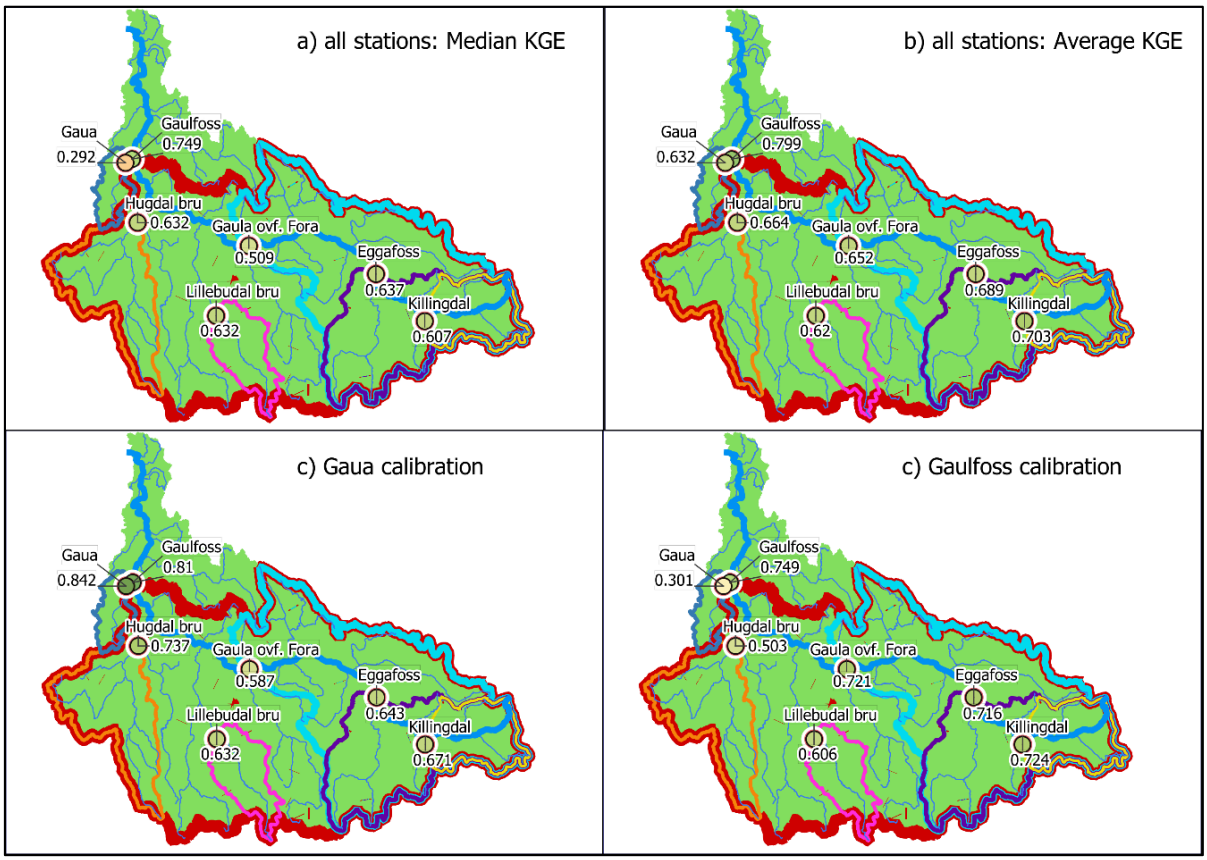


Figure 31: KGE of parameter set resulting from calibration with median KGE (upper left), average KGE (upper right), Gaua (lower left) and Gaulfoss (lower right)

5.2.1.4 Low flow calibration

The low flow calibration was done to test the influence of input data towards poor low flow performances. Different approaches and their results are described in Appendix E.2. After testing everything, no promising solution was found. However, seNorge2 had a better performance of the low flows in Nausta and Usma, compared to NWP.

5.2.2 Soil layer calibration

To check whether another amount of soil layers or soil thicknesses will improve the model performance, these values were varied in `GeoClass.txt`. and the model calibrated. In this calibration, the progressive Monte Carlo method with 10 x 200 simulations was applied using the seNorge2 dataset. The resulting KGE of each catchment with different soil layer properties is summarized in Table 12. Most of the catchments have the highest performance with three soil layers. The only exception is Grunnåi, that performed best with two soil layers.

Table 12: Results of soil layer calibration in all of the catchments.

		Nausta	Usma	Gaula	Surna	Grunnåi
n layers	**Soil thickness [m]**	**KGE**	**KGE**	**KGE***	**KGE**	**KGE**
1	**1**	0.852	0.665	0.588	0.657	0.684
	2	0.864	0.620	0.423	0.669	0.636
2	**1**	0.855	0.651	0.601	0.648	0.642
	2	0.841	0.680	0.607	0.657	0.688
	3	0.874	0.701	0.579	0.672	0.590
3	**1**	0.871	0.604	0.580	0.655	0.677
	2	0.911	0.706	0.586	0.674	0.509
	3	0.901	0.718	0.647	0.662	0.650

* Average KGE of all catchments

After calibrating the soil thickness, the best results were further analysed by varying the stream-depth. For that, it was set on ground level (0) and 25 cm above and beneath it. The same calibration method as in the soil layer calibration was applied. The results are summarized in Table 13. Every catchment had the best performance for a stream-depth similar to the ground of the soil, meaning that there is no standing water in the catchment in that case.

Table 13: Results when changing the stream-depth with the best results from Table 12

Stream depth Difference to layer ground	**Nausta**	**Usma**	**Gaula**	**Surna**	**Grunnåi**
	KGE	**KGE**	**KGE**	**KGE**	**KGE**
-0.25 m	0.911	0.872	0.637	0.651	0.531
0 m	0.924	0.877	0.647	0.688	0.733
+0.25 m	0.856	0.829	0.601	0.680	0.661

All of these calibrations were done with seNorge2. To test if there is a dependence on the dataset, the stream depth was varied in Nausta and the NWP dataset. Here, the highest KGE was calculated with a stream-depth of 25 cm above the soil ground.

5.3 Model results

After finishing the catchment calibration, the model performance was compared, characteristics of the modelling results were analysed more in detail and simulated discharge from each dataset was compared to the actual observation. In the next step, water balances were created for each catchment for an annual and monthly timestep. Storage changes in the model were analysed additionally. Finally, specific events were portrayed and evaluated by comparing model results, precipitation data and precipitation calculated from an inverse modelling approach.

5.3.1 Performances

After setting the soil layers to the ones with the best results of the soil layer calibration, the main calibration process could start. For that, the algorithm from chapter 4.5.2 was applied. The final performance of the models between 09/2014 and 09/2016 is summarized in Figure 32. The performance of each catchment is given when using seNorge2, NWP and weather station data (WS) as input into the model. Furthermore, each catchment was one time calibrated using precipitation correction factors *pcurain* and *pcusnow* and one time without them. The validation is included as well. The weather station of Grunnåi is an exception because it was calculated with daily values (instead of hourly). Usma was not validated because of a lack of long enough discharge time series.

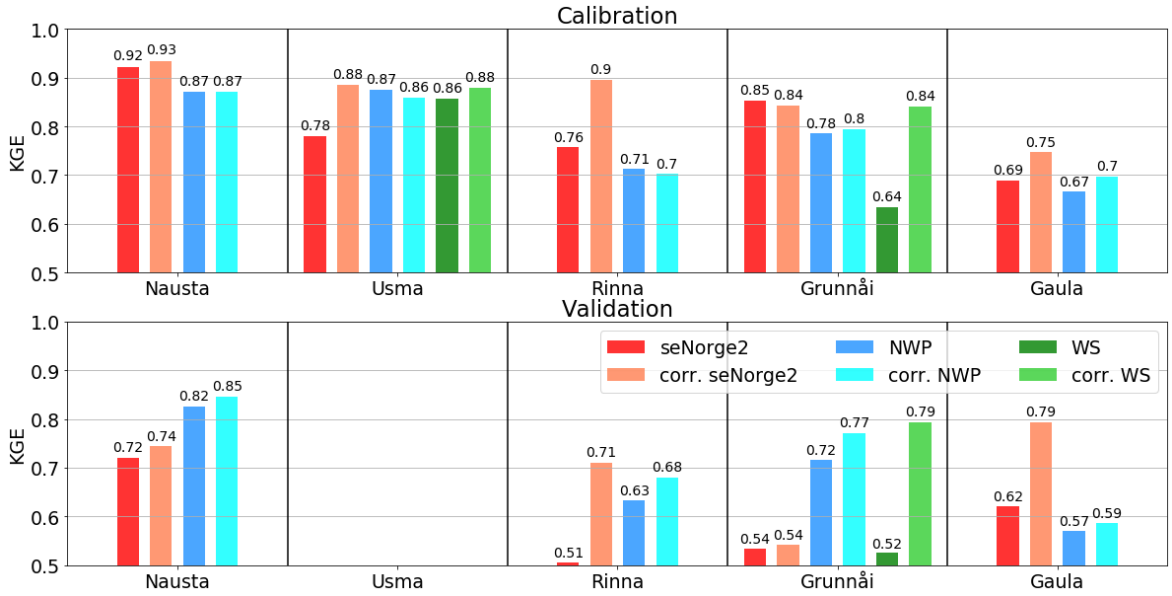


Figure 32: Model performances of each catchment using different input datasets

Gaula is divided into different sub-catchments. The results are summarized in Figure 33. All of these performances are based on the optimization of the average KGE. The results appear to be variable and seNorge2 and NWP performances vary in each sub-catchment. The KGE of the validation seems to often increase when using seNorge2 as input and is decreasing when using NWP.

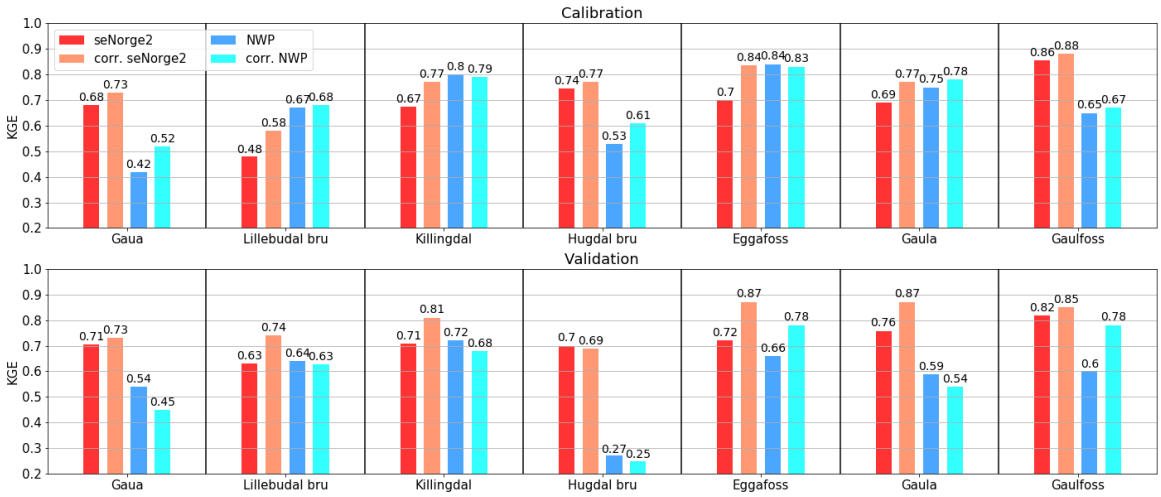


Figure 33: Performance of each catchment in Gaula with different datasets

5.3.2 Individual catchment results

To get a more detailed information about the model behaviour within each catchment, the results were evaluated by their Cumulative Distribution Function (CDF), correlation towards observations, performance in summer and winter and improvement after the stepwise calibration. In this chapter, an example of the analysis process is given in the first part, afterwards the results of each catchment are briefly summarized.

The detailed result of Nausta using seNorge2 with snow and rain correction as input is described in the following. First, the correction parameters *pcurain* and *pcusnow* are read out of the resulting parameter set from the calibration. In this case rain is increased by a constant factor (*pcurain*) by 5 % and snow (*pcusnow*) by 17 %. Afterwards, the improvement of the KGE and NSE during the stepwise calibration is calculated. Here, the KGE increased by 0.06 compared to the initial calibration with 25,000 simulations, when fine-tuning the parameters. The KGE of the final model is 0.934 with a slightly lower performance in summer. When looking at the CDF (see Figure 34) low flows are in good agreement with observations, but are slightly underestimated. Peaks are underestimated as well. The CDF was separated into winter (October to May) and summer (May to October) and analysed in a similar way. In summer, peaks and low flows show a good agreement with the observations. In winter, peaks and low flows are underestimated.

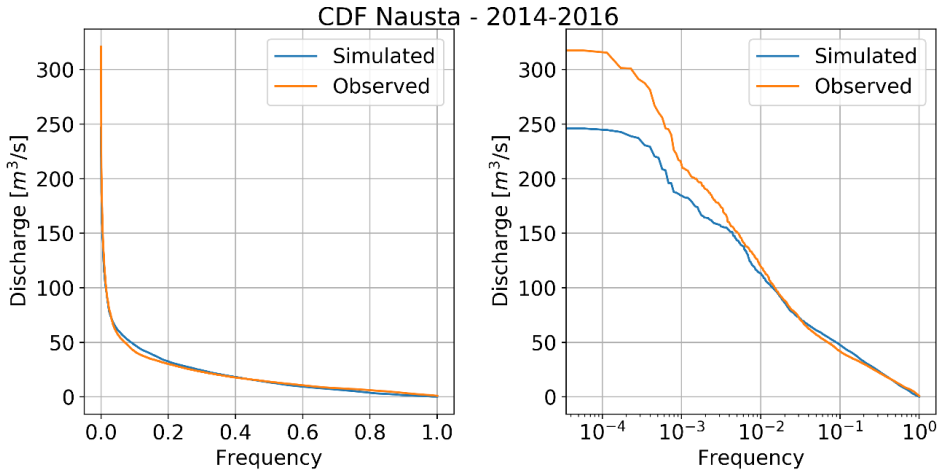


Figure 34: CDF Nausta modelled with SeNorge2 and precipitation correction

Furthermore, a comparison of observed and simulated discharge was done by scattering these values (see Figure 35) and calculating the correlation (r^2) in summer and winter. In this case, winter values have a better correlation to observations. The linear regression shows that the simulated discharge is slightly underestimating the discharge. This is just an example of the analysing process. Other figures can be found in the digital appendix.

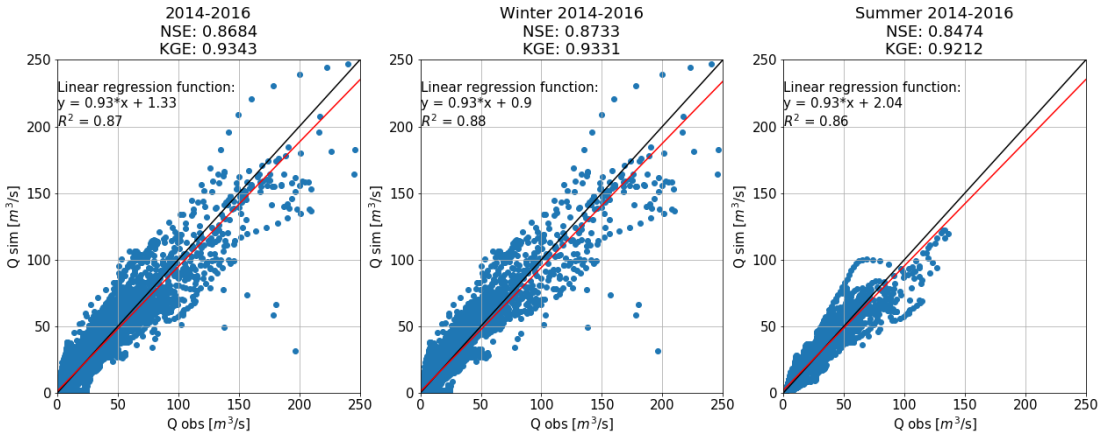


Figure 35: Scatter Nausta modelled with SeNorge2 and precipitation correction

The analysis explained above is summarized for Nausta in Table 14. Besides the corrected version of seNorge2, it also contains the results of uncorrected seNorge2 and NWP. Here, the performance is slightly worse. All of them perform poorer in summer, the difference is higher with NWP. NWP seems to underestimate low flows more than seNorge2. The correlation is also lower when using NWP (corrected and uncorrected) as model input.

Table 14: Summary results Nausta

		SeNorge2	SeNorge2 corr	NWP	NWP corr
Correction	**pcurain**	-	0.05	-	0.004
	pcusnow	-	0.17	-	-0.083
Difference calibrations	**KGE improvement**	0.003	0.006	0.003	0.003
	NSE improvement	0.009	0.001	0.008	0.008
Performance	**KGE**	0.922	0.934	0.871	0.871
	KGE summer	0.916	0.921	0.779	0.781
	KGE winter	0.926	0.933	0.885	0.880
	NSE	0.864	0.868	0.756	0.760
	NSE summer	0.865	0.8474	0.689	0.733
	NSE winter	0.856	0.873	0.774	0.765
CDF	**Low flows***	(su)	(su)	(u)	(u)
	Low flows summer*	(u)	(g)	(u)	(u)
	Low flows winter*	(u)	(u)	(u)	(u)
	Peaks*	(u)	(u)	(u)	(u)
	Peaks summer*	(u)	(su)	(o)	(g)
	Peaks winter*	(u)	(u)	(u)	(u)
Scatter	**r^2 all**	0.87	0.87	0.78	0.78
	r^2 summer	0.87	0.88	0.79	0.82
	r^2 winter	0.87	0.86	0.78	0.78

* underestimated (u), slightly underestimated (su), in good agreement (g), slightly overestimated (so) and overestimated (o)

The results of the other catchments can be found in Appendix F and the digital appendix. Here, they are briefly summarized. In Usma, the uncorrected precipitation set of NWP yields to the best model performance and correlation compared to observations. It also leads to a better performance in winter. When using correction factors for NWP, the model performance is not improved. The corrected model with seNorge2 leads to better results and correlations. Usma has a poorer agreement in winter than in summer for all datasets. The weather station does not create the best results but has however the best fitting low flow conditions. In Grunnåi, seNorge2 yields to better performances than NWP and the weather station. The peaks are better estimated and the correlation is higher as well. All datasets lead to a low performance in winter especially with NWP and the uncorrected weather station data. SeNorge2 is calibrated with a better performance in Surna as well (corrected and uncorrected). However, NWP has a better correlation with observed values, especially in winter. Both datasets lead to an underestimation of low flows. In Gaula, all catchments underestimate low flows and dry out regularly. The KGE is always worse in winter than in summer. There are catchments with a better performance of seNorge2 and some with NWP data. The same happens with the correlation. The corrected version of the datasets does not lead to an improved performance in all catchments.

To have a closer look on simulated discharges, the cumulative volume sum of each dataset was compared within each catchment. The spread of simulated discharge was visualized by filling the minimum and maximum discharge estimated with the datasets. The results of Nausta are visualized in Figure 36. Figures of other catchments can be found in Appendix F. In Nausta, the spread between discharges is not high. In general, all observed data is in between the model results. However, there are some events were no dataset agreed with the observations. For example, in January 2016, where no model calculated the discharge that was observed. In October 2014 as well as April 2016 assumed the datasets a discharge peak, that was not observed. Additionally, the model underestimates low flows with every dataset in winter. The cumulative volume sums show that seNorge2 is below the observed volume, the corrected version is however more or less similar and additionally has the same slope as the curve. NWP is above the curve, the corrected version even higher.

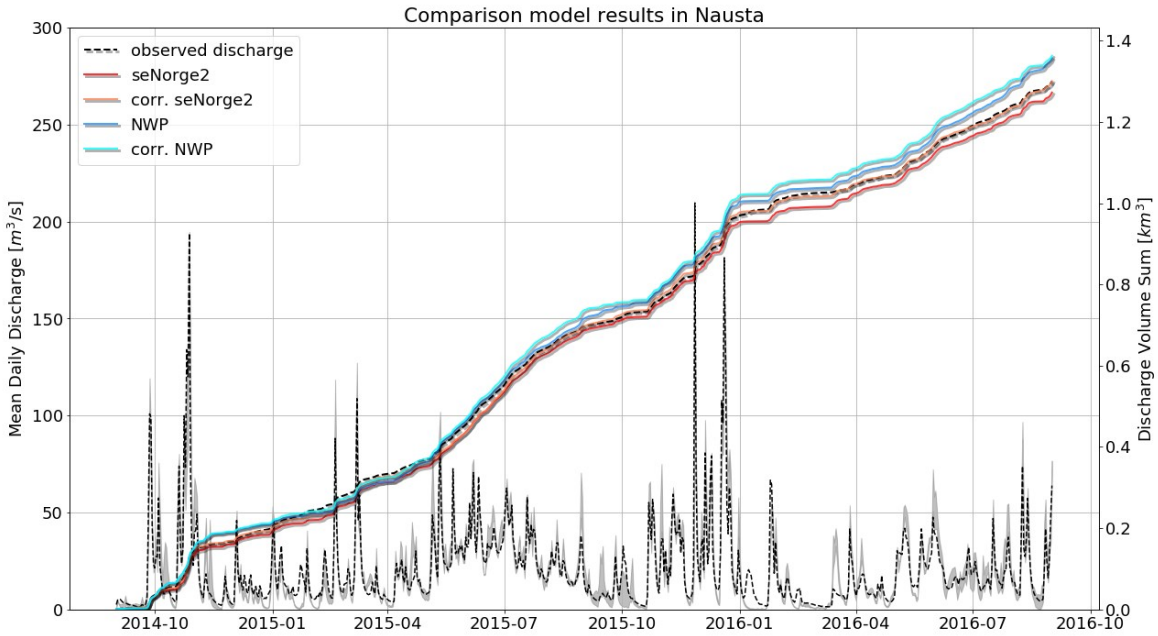


Figure 36: Comparison modelling results of Nausta using different input datasets. The time series was averaged to a daily timestep. Grey area: simulated discharge from the different products

When looking at the results of other catchments, the spread of the simulated discharges of the datasets is bigger compared to the observed discharge in Usma and Grunnåi. A low flow underestimation of all the datasets is found in Rinna and Grunnåi as well. In Usma, there are events where no dataset lead to a discharge peak, for example in October 2014 and March 2016 and on the other side where models wrongly assumed discharge peaks, for example in July 2016. In Grunnåi, there are less events where discharge was not simulated at all from the datasets. However, in June 2016 the spread of one event was particularly high. An inverse example, where a discharge peak is assumed by every dataset but was not measured is March 2016. Rinna has some events where some datasets didn't show precipitation, for example September 2014 and December 2015. In April 2015 and in June 2015 the models simulated a discharge peak that was not observed by the discharge gauge.
Gaula is divided into seven different hydrographs for each gauging catchment. These results vary strongly. In the catchments Gaua, Hugdal bru and Killingdal are a lot of fast increasing and variable discharge peaks, whereas Gaulfoss is smoother. The spread of the models with different datasets is wide, but there are not many events, where the behaviour differs strongly. Some events were found that had similar results as the ones described in the last chapter. They will be portrayed in detail in chapter 5.3.4.
The cumulative discharge volume modelled with seNorge2 is always below the observations and also the smallest one of the datasets, except for Grunnåi, where the model with weather station data simulates an even smaller discharge volume and Gaua, where NWP values are smaller. The model with model correction factors is volume wise close to the observations in Rinna and Usma. NWP has a good agreement in Usma and Grunnåi, but is highly overestimating the discharge in Rinna. NWP with precipitation correction leads to an overestimation in all of the catchments, except for Gaula.

5.3.3 Water balance

For further analysis, the modelling results were evaluated with the water balance of each catchment. First, the annual water balance during the study period was portrayed. Then seasonal behaviour was analysed for each catchment in detail. To make sure there is no water accumulation in a storage, the water volume of different storages (e.g. lakes, groundwater) was controlled. In the last subsection of this chapter, PET from MOD16 was compared to the one calculated with Hargreaves.

5.3.3.1 Annual water balance

All sub-basins were averaged according to their area to calculate the water volume. To get comparable results of the datasets, only simulated values between 2014-2016 (September to September) are included. The annual water amount of both years was then averaged, and the resulting values visualized. In Figure 37 are the results of all catchments except for Gaula. The results of sub-catchments in Gaula are in Appendix F.1.
The figure visualizes the different water volumes in the catchment. In general, the precipitation amount varies the most. SeNorge2 and the weather stations usually assume the lowest amount of water. The evapotranspiration varies as well and is assumed to be lowest when using seNorge2 and weather stations data. The black line represents the evapotranspiration amount estimated by MOD16 and can be used for comparison. In general, it assumes the highest amount of evapotranspiration in most of the catchments compared to the model. The discharge volumes are close to each other for each dataset, and close to the observed discharge with some exceptions. The annual storage change simulated by the model is in most cases close to zero. However, NWP in Usma leads to an accumulation of storage, seNorge2 to a decrease. In Nausta, there is an increase

in the water storages as well for all of the datasets. The storage changes calculated with the precipitation products, simulated evapotranspiration and observed discharge data show more variation in the volume. They agree with the under- or overestimation of discharges with precipitation of the datasets. They display the water storage change in the catchment when precipitation data is assumed to be true, and the modelled evapotranspiration is correct.

Gaula (see Figure 85) includes different catchments, but the correction factors are calibrated equally for each of them. The model with seNorge2 always leads to an underestimation of the discharge and low evapotranspiration amounts. The corrected version increases the evapotranspiration and discharge, but is still, except for Gaulfoss, lower than the observational data. Using NWP leads to an overestimation in Gaula ovf Fora. Whereas the discharge volume in other catchments is underestimated. This also leads to varying influence of the correction factor in these catchments. Evapotranspiration by MOD16 is higher than most of the model results. Only in Gaua, NWP assumes a slightly higher AET than estimated by MOD16. Most catchments have barely any storage changes within the model.

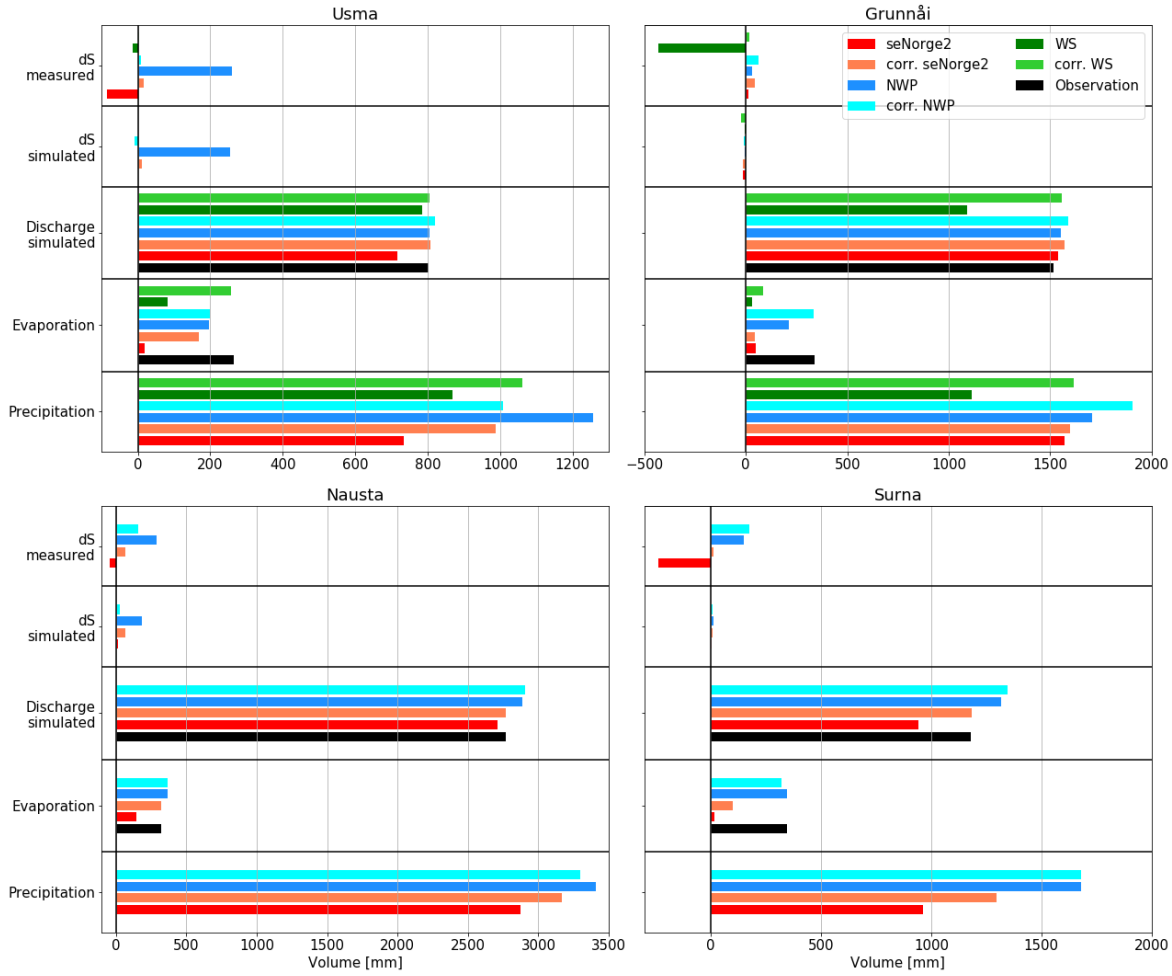


Figure 37: Annual water balance of the catchments Grunnåi, Nausta, Surna and Usma. Values are averaged between 2014 and 2016. Observations are discharge measurements and MOD16 evapotranspiration data

5.3.3.2 Seasonal cycle

To further investigate the influence of these datasets on the water balance, the seasonal cycle of precipitation, evapotranspiration, discharge and water changes were analysed. For that, the monthly precipitation, evapotranspiration and discharge amount was accumulated, and the storage changes were calculated using the water balance equation. This was done for every catchment individually.

Figure 38 shows an example of Usma. Here, precipitation (first row) is divided into month from September to August. The values are an average between 09/2014 and 09/2016. The monthly precipitation sum is then compared between seNorge2, NWP and weather station data. There is a lot of variation between these volumes. There is no month where seNorge2 estimates more precipitation than NWP. The seasonal behaviour is similar within all datasets with less precipitation in winter than in summer. The corrected versions of NWP and the weather station vary stronger from the original in summer than in winter.

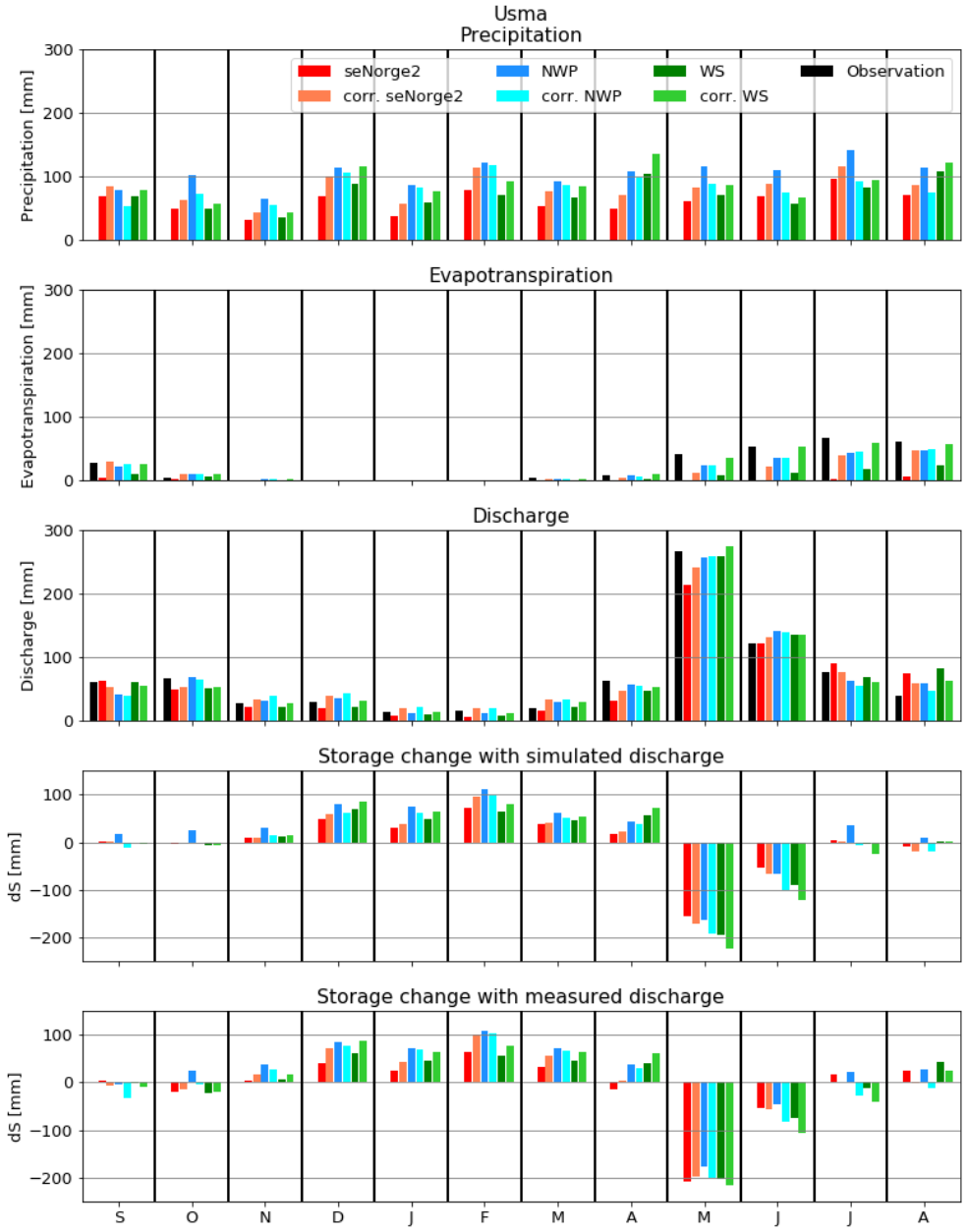


Figure 38: Seasonal water balance of Usma when using different input products. Observed values are from discharge measurements and MOD16.

The second row shows the seasonal evapotranspiration cycle. Here, MOD16 data is included as a reference. Evapotranspiration in Usma only occurs from Mai to October. The model with seNorge2 barely assumes any AET, more when corrected. For most cases, MOD16 assumes more evapotranspiration than the model estimates. The monthly discharge volume shows another relation between the products. In general, the volumes vary less. From June to September, seNorge2 creates a higher amount of discharge than NWP. The snow melt flood is mostly in May and June. Here, NWP shows a better agreement with the discharge amount. The storage changes are (except for NWP) almost zero from July to October. Between November and April, the storage is increasing, in Mai and June, the storages decrease. Simulated and measured storage changes are mostly similar, except for summer with higher amounts of the measured storage.
Looking at the other catchments (see Appendix F.1) shows similar results. In Nausta, seNorge2 precipitation is lower than NWP, but the corrected version is almost similar to it. Evapotranspiration with seNorge2 is significantly lower than MOD16, NWP is a bit higher. There is no significant snow melt flood in Nausta during spring. NWP tends to overestimate discharge in summer. The storages are increasing from October to April and decreasing in the summer month.
Grunnåi is an exception to other catchments, because here, seNorge2 estimates a higher amount of precipitation in September and August compared to NWP and weather stations. Compared to NWP, the weather station data is especially low in the winter. The corrected versions of this catchment are close in term of precipitation amount in each month. The evapotranspiration is always assumed to be lower than MOD16. Only the corrected version of NWP is close to MOD16. All datasets lead to lower discharge volumes in winter, NWP leads to an overestimating the discharge in May and June (snow melt). In terms of storage change, it was found that models using NWP have higher storage increases in winter and decreases in spring.
In Surna, seNorge2 is again estimating the lowest amount of precipitation in every month, the corrected version of NWP is on contrary high compared to seNorge2 in winter. Evapotranspiration is again low for seNorge2, NWP estimates are close to MOD16. SeNorge2 leads to better agreement of simulated and observed discharge volume except for winter and the spring snow melt flood.
Figures of the seven catchments of Gaula can be found in the digital appendix. In Gaulfoss and Hugdal bru, seNorge2 and NWP mostly differ in summer, but is almost similar in winter. Same applies for the discharge. In Eggafoss, Killingdal and Lillebudal, NWP is higher than seNorge2 in every month. The spring flood in May is underestimated strongly using both datasets in Killingdal, but is in good agreement with NWP in Lillebudal. High discharges in January are not simulated by any model in Lillebudal and Hugdal. In Gaua, seNorge2 estimates higher values than NWP, except in summer, leading to higher simulated discharges.

5.3.3.3 Storage changes

A last step to compare the model' water balances was to assess storage changes within the model. The annual water balance already showed that there is no accumulation in most of the catchments, but in some cases, there was an increase or decrease of storages. That's why all the water balance files were analysed and visualized over time to see whether there is an accumulation or decrease of water volume in one of the storages.
The water balance was analysed and the storages changes were calculated by summing up the storage volumes. To test whether all water storages are included, the storage changes were calculated as well using the water balance equation (Eq. 1). They were the same in all catchments and only varied in a small manner (up to 3 mm in Usma, but usually below 1 mm of difference).
These figures of all catchments are in the digital appendix. Appendix F.1 includes three examples of the storage behaviours that were found in the catchments. Figure 89 shows a catchment with no storage changes in the model. During winter, there is an increase of storages through snow, that decreases in spring. That behaviour was seen for most of the model results.

However, there are model results, where the storage increases. Figure 90 shows an example of Usma. There, the storage increases over the two years, by an accumulation of water inside the river. The same process happened for the corrected seNorge2 version in Usma. This model is however less strong accumulating. Figure 91 shows an example of snow accumulation in Nausta where snow melt is not fast enough to melt snow in the upper catchments before the next winter. This is the case for the NWP dataset as well as the corrected version of seNorge2.

5.3.3.4 Comparison evapotranspiration

To evaluate the calculated PET, it was compared with MOD16 data of each catchment. Figure 39 shows a timeseries of MOD16 PET and AET compared to PET calculated with Hargreaves in Gaula. It shows that Hargreaves values are in between PET and AET from MOD16, meaning that the calculated PET is lower than the one given from MOD16. Still, it is never lower than AET. There are no values of MOD16 in the winter month, probably because it is assumed to be no evapotranspiration then. Hargreaves is close to zero in these months.
The evapotranspiration time series of other catchments can be found in the digital appendix. They behave in a similar way, except Hargreaves PET of Grunnåi is close to MOD16 AET, but still almost always higher.

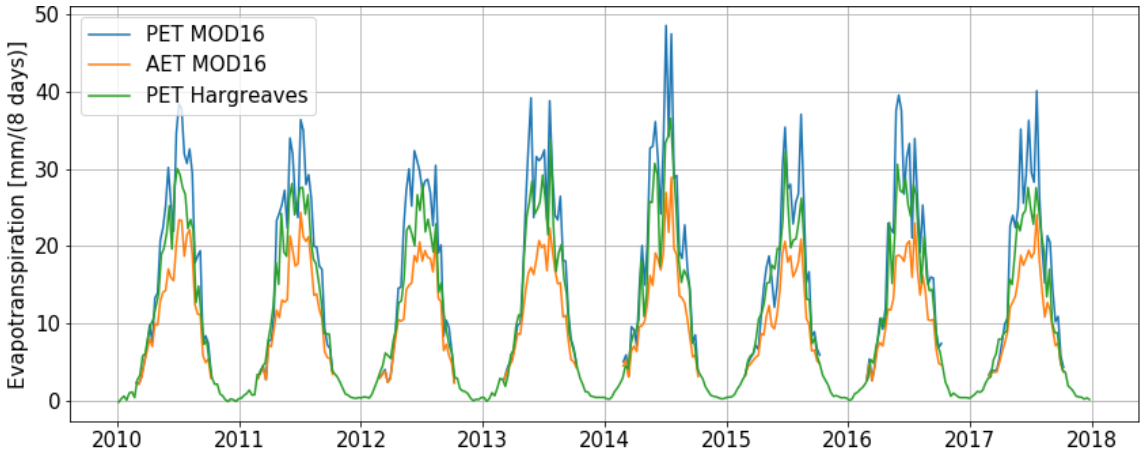


Figure 39: Evapotranspiration from Hargreaves and MOD16 over time (2000-2018) of the Gaula catchment

5.3.4 Analysis of selected events

After looking at model performances and water balances, the datasets were further evaluated by looking at specific events. Figure 40 shows an example of such an event. The events were further visualized as a video of the spatial distribution of precipitation over time. These videos can be found in the digital appendix.
The example shown here is in Usma on the 13th of July in 2015. Here, a discharge peak was measured from the discharge gauge, but no model assumed this behaviour of the catchment. The weather station and seNorge2 didn't estimate any or just a small amount of precipitation at that time. NWP estimated a precipitation amount of 2.5 mm in one hour.

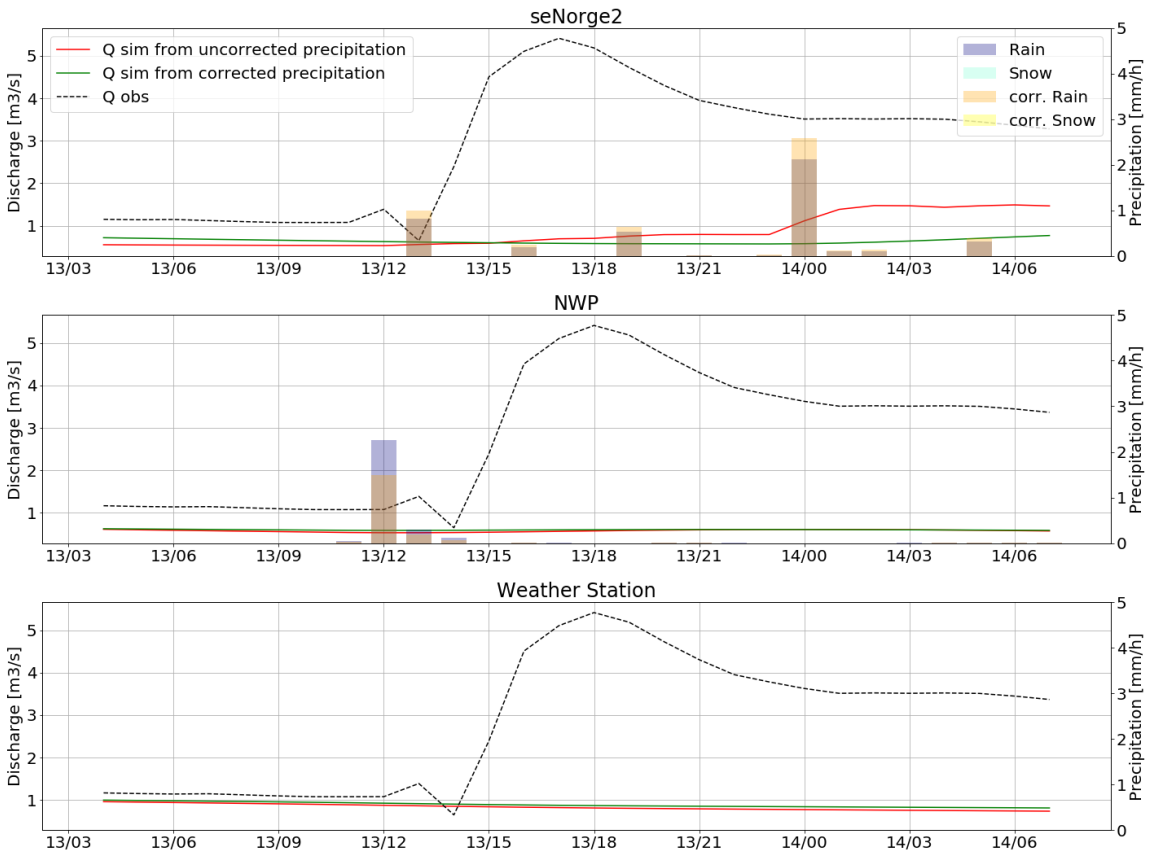


Figure 40: Precipitation event in Usma, July 2015

To test, how much precipitation should have been estimated according to the discharge data, the method of Kirchner (2009) was applied by using an inverse model for estimating the precipitation amount. The results can be seen in Figure 41. The inverse model predicts that there have been two precipitation events. One at 12 am with 2 mm and one at 2 pm with 8 mm. NWP estimated the same precipitation amount at 12 am, however missed the other one in the afternoon. SeNorge2 missed the first event and highly underestimates the second one.

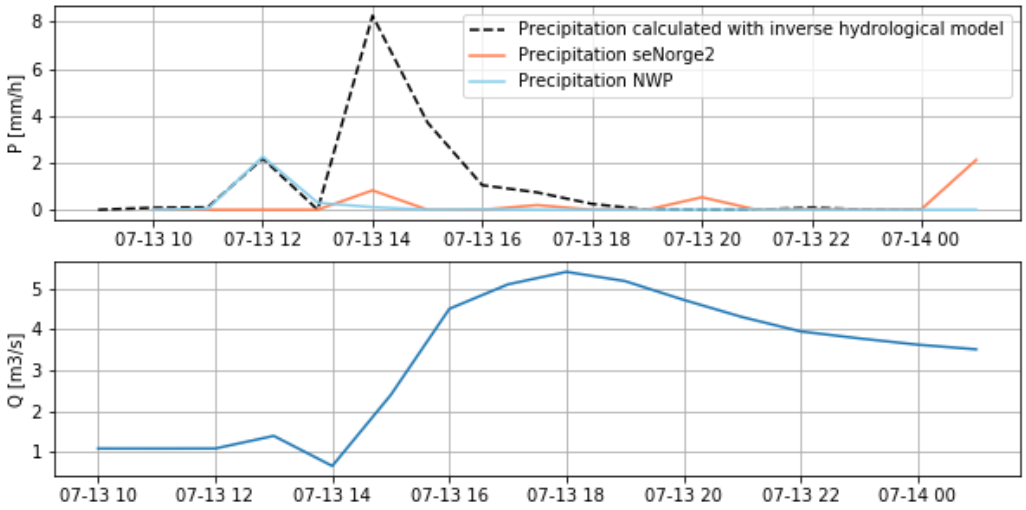


Figure 41: Precipitation estimated from seNorge2, NWP and an inverse hydrological model in Usma, July 2015

More events were looked at and can be found in Appendix G. Usually, when the model did not react to discharge that was observed by the gauge, precipitation was not measured enough or not at all. This behaviour can be seen in the event in Usma, October 2014 (see Figure 92), where seNorge2 did barely assume any precipitation, NWP about 5 mm, but the model showed no major outflow response and the weather station measured about 7 mm leading to a good correlation between simulation and observation discharge. The inverse hydrological model assumed a precipitation sum of 20 mm (see Figure 93). When looking at the spatial variability of the event (see digital appendix) the precipitation is locally concentrated.
Another similar event to this is in September 2014 in Rinna (see Figure 94), where the model did not react to precipitation from NWP, but was in better agreement with seNorge2. In this case however, both datasets assumed the same precipitation amount. The uncorrected seNorge2 model is overestimating the discharge, although less precipitation is assumed compared to the corrected precipitation amount. A similar behaviour was found in Eggafoss in October 2014. SeNorge2 leads to a good flood response with the precipitation amount, whereas the model calibrated with NWP did not react at all.
In some cases, missed discharge peaks can also be caused by temperature. One example is in Nausta in January 2016 (see Figure 95). Here, temperature data is included. No dataset assumed heavy precipitation, nevertheless the discharge increased during that time. Temperatures were at about -5 °C and only increased during a snow event five days after the original observed discharge increase. Another example is in Rinna in December 2015, where a discharge peak was measured, but not simulated, because temperatures were below -5 °C and only later rose above 0 °C, leading to a delayed discharge peak (see Figure 96). Another example is in Hugdal bru (Gaula) in January 2016 (Figure 101), where a high discharge was measured even though there was no precipitation assumed and temperatures were below -10 °C.
Then there are events where the model calculated discharge that was not measured by the gauge. One example of this phenomenon is in June 2016 in Grunnåi (see Figure 97), where seNorge2 estimated a high precipitation amount leading to a discharge peak, that was not measured by the discharge gauge. NWP estimated only a small rain rate, leading to no particular increase of discharge in the model and is therefore in better agreement with the observation. Another good example is in Surna, June 2015 (see Figure 98). Here, only NWP estimates precipitation, whereas seNorge2 didn't estimate any. In Gaua (Gaula) in July 2016 NWP estimates higher precipitation rates leading to a flood response that was not measured by the model.

5.4 Sensitivity analysis

In the last part of the study, the sensitivity of the model and the results were evaluated. For that, GLUE was applied, the calibration and validation period switched, and physical parameter compared and analysed.

5.4.1 GLUE

GLUE analysis was applied for two catchments. At first, the parameters were normalized and compared with each other. These results can be found in Appendix H, Figure 102 and Figure 103. There are mostly no parameters converging against the boundaries. In Grunnåi, some parameters like *pcusnow* and *rivvel* are close to the boundary. There are some parameters that don't have varying values (e.g. *rivvel*, *cmelt* and *pcurain*), whereas others spread a lot (e.g. *surfmem*, *wcfc*, *wcep* and *fcdistmax*).The same results were found in Nausta, except that there was a wider spread of *rivvel*, *cmelt* and *pcurain.* Furthermore, the river velocity showed different values depending on the input dataset.

To get a more detailed look into the parameter results of each run, the values were visualized in more detail. This is shown in Figure 104 in Appendix H. There are parameters, that don't vary between each model run like *rivvel, pcurain, pcusnow, damp*. Then there are parameters, that vary with each model run, but not within the best five results (like *cmelt, grata, gratp, ttpd*). There are also parameters varying in each run and within the best five results (e.g. *cevpph, cevp, srrcs, fsceff*). In general, there is a wide spread of parameter space for some of the parameters. An example of the three different types is shown in Figure 42 with *rivvel*, *grata* and *cevpph*.
Figure 105 shows the same analysis for Nausta. There are less parameters with no spread between each run. All parameters that did not change in each run in Grunnåi, vary in Nausta (*rivvel, pcurain, pcusnow, damp).* Other parameters (like *cmelt, grata, gratp, ttpd*) however behave similarly.

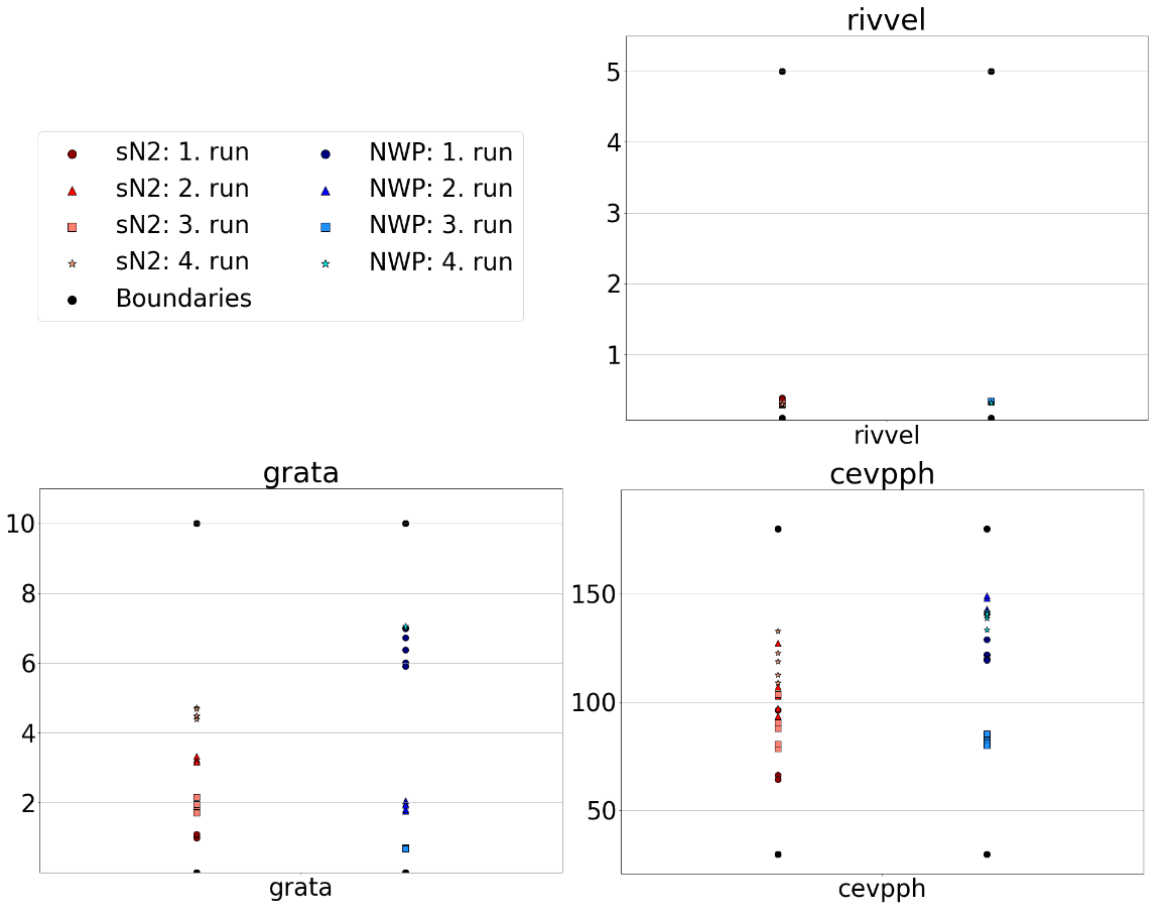


Figure 42: Example of the calibrated parameter space of *rivvel*, *grata* and *cevpph* in Grunnåi

5.4.2 Variating calibration and validation period

Additional to GLUE, a more simplistic approach was done, by switching calibration and validation period. This was done for three catchments: Nausta, Grunnåi and Rinna. The results are summarized in Table 15. To understand the table structure, an example is given here. In Nausta, seNorge2 was originally calibrated between 2013/09 and 2016/09 (time period 1) resulting in a KGE of 0.918. The validation was between 2011/09 and 2013/09 (time period 2) and resulted in a KGE of 0.725. When switching the calibration period to time period 2, the KGE is 0.913. Validation in time period 1 leads to a KGE of 0.924.

Table 15: Results of analysing the influence of a switched time period of calibration and validation

	Time period 1		Time period 2	
	Calibration	Validation	Validation	Calibration
Nausta – SeNorge2	0.918	0.924	0.725	0.913
Nausta – NWP	0.879	0.748	0.844	0.861
Grunnåi – SeNorge2	0.845	0.367	0.507	0.825
Grunnåi - NWP	0.809	0.811	0.759	0.807
Rinna – SeNorge2	0.866	0.704	0.703	0.790
Rinna – NWP	0.824	0.583	0.674	0.778

These results were further analysed by plotting a double sum curve of precipitation and simulated discharge data for Rinna with both datasets (see Figure 107, Appendix H). It was found that the assumed precipitation amount (through correction factors of the models) varied depending on the chosen time period of the calibration.

5.4.3 Analysis of physical parameters

In the last step of this analysis, specific parameters were compared considering their physicality as well as their value with different datasets. At first, the parameters *pcusnow* and *pcurain* were evaluated in each catchment and for seNorge2 and NWP data. The values vary within the catchments, especially NWP, but even more between the two datasets. The results are visualized in Figure 43.

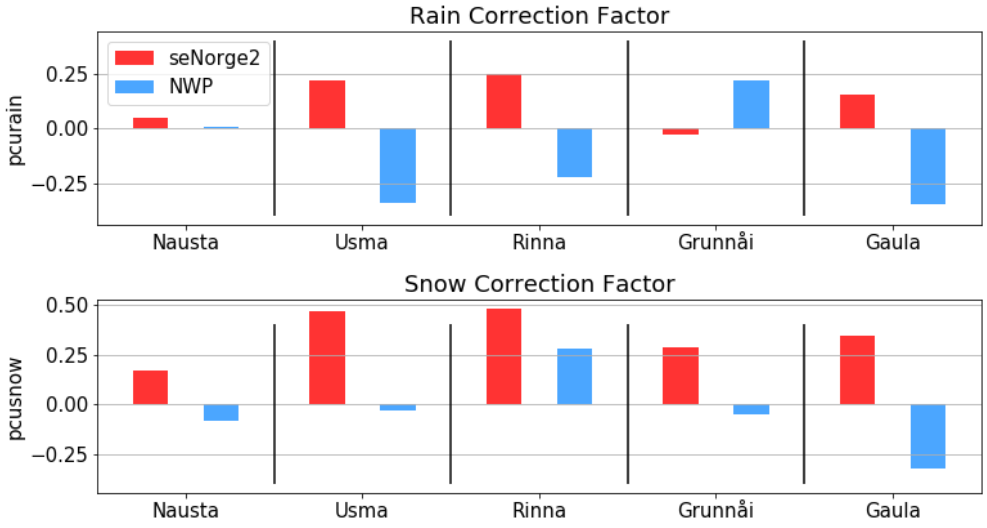


Figure 43: Variation of *pcurain* and *pcusnow* in each catchment and for seNorge2 and NWP

Another parameter that was analysed is the river velocity (*rivvel*), because it is a physical measurable value. The results are shown in Figure 44. This comparison shows the variability of this parameter in each catchment, but also with each dataset. To further asses this finding, the river velocity was compared with the mean slope and elevation of the catchment. This is shown in Figure 106, Appendix H. There was however not a clear correlation visible between these values.

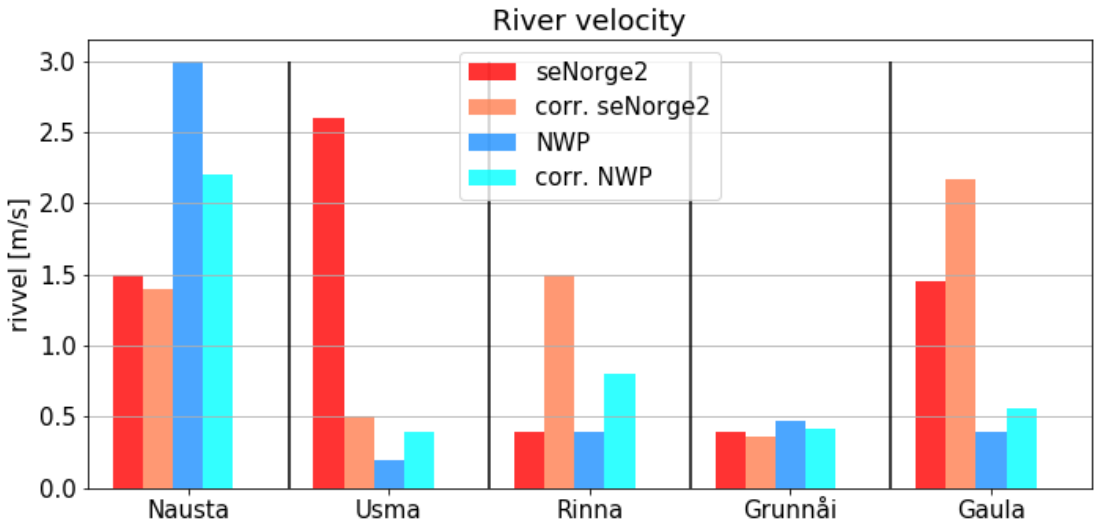


Figure 44: Comparison parameter rivvel in the catchment and with seNorge2 and NWP

In the last step, a SLC dependent parameter was compared. The snow melt factor *cmelt* is calibrated separately for each SLC. The resulting parameter values are shown in Figure 45. Although they vary in general, there are some common behaviours. For example, parameters are often not changing between seNorge2 and NWP (with some exceptions). Often, the snow melt factor seems to be lower in forest and marsh areas, whereas mountains and water are a bit higher in most catchments. In Usma and Rinna however, the snow melt factor of mountainous and forest SLCs is the highest compared to other values.

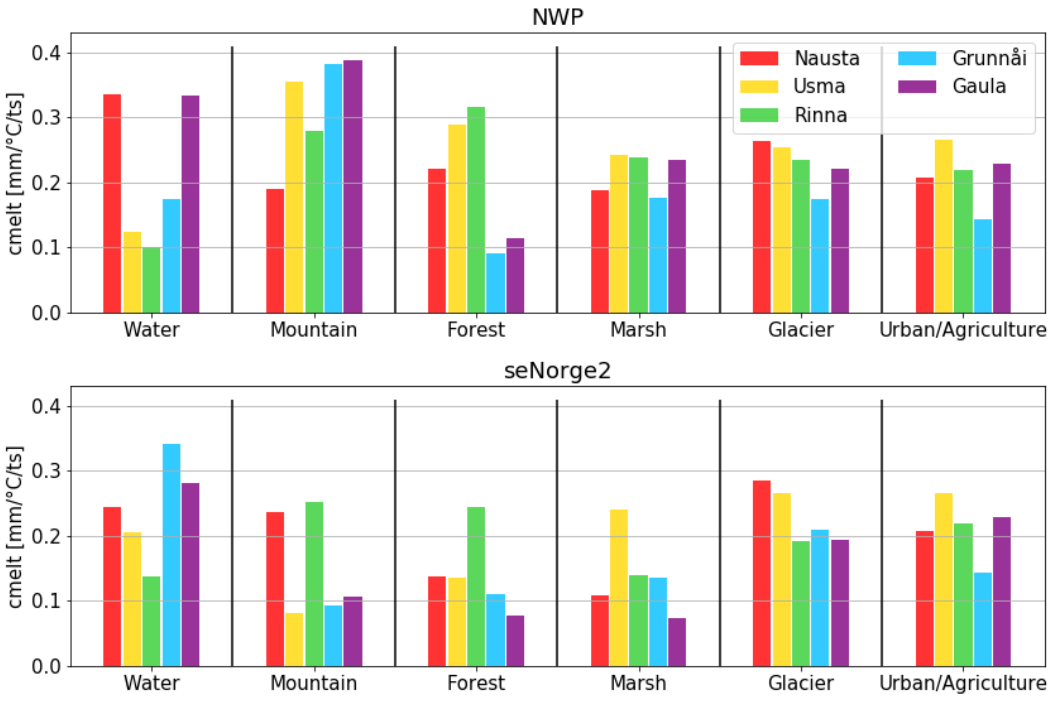


Figure 45: Snow melt factor cmelt for each SLC and catchment. Upper part: NWP, lower part: seNorge2

6 Discussion

In this chapter, the results will be discussed and analysed. At first, the comparison between datasets will was evaluated, then the model results compared, by looking at the performance, water balance and specific events. At last, the sensitivity analysis will be discussed.

6.1 General dataset comparison

In a first step, the general behaviour of precipitation products was analysed and compared. The NWP product is based on radar and gauge data, but also includes a numerical weather product as background. This leads to fewer errors, because for example radar beam blockage can be balanced by the numerical model, which in return is also based on radar data (and other observations) as an initial state. Annual precipitation sums revealed that NWP always estimates the highest precipitation amount in each of the catchments. One reason for that might be the wind correction of gauge data, causing higher precipitation amounts, especially for snow. The annual precipitation sum was further analysed by using discharge measurements and MOD16 evapotranspiration to calculate the water balance. NWP seems to under- and overestimate precipitation depending on the year and catchment. In Usma, it is overestimating the precipitation amount, which is in agreement with the comparison of NWP and a weather station near Usma using daily precipitation sums. However, this might also be caused by gauge measurement errors, for example by wind-induced under catch of precipitation.

SeNorge2 always estimated a lower amount of annual precipitation in each catchment. Looking at the water balance revealed that seNorge2 is underestimating annual precipitation in every study catchment by up to 900 mm/a. The underestimation is especially high in Nausta, Grunnåi and Rinna. One reason for that might be the high elevation of these catchments and the knowledge, that seNorge2 tends to underestimate precipitation in mountainous regions (Lussana et al. 2018). Another reason might be that wind-induced under catch is not corrected, causing an underestimation especially in winter. Compared to weather station data, the underestimation was also detected when looking at the linear regression of daily precipitation data.

Looking at the spatial variability of annual precipitation sums revealed the influence of the DEM background field of the interpolation. For example in Grunnåi (Figure 60), lower elevated areas have lower precipitation sums than higher elevated areas. In Nausta (Figure 59) it is the other way around. Even so precipitation can be dependent on the elevation (for example in Nausta because of orographic rain that is caused by humid air from the sea rising at the mountains and cooling down causing water to condensate), however there are limits to this assumption, for example by looking at small valleys.

Regarding the water balance analysis, it needs to be considered that MOD16 is known to overestimate evapotranspiration (Huang et al. 2019), which would lead to the assumption that precipitation is more underestimated than it actually is. Still, the error of evapotranspiration can be assumed to be smaller than the one from precipitation, because the amount of evaporated water is much smaller than precipitation.

Evaluating radar data revealed that it measures the lowest amount of precipitation compared to the other dataset. Because seNorge2 is already underestimating the precipitation amount, it can be concluded that radar is highly underestimating precipitation. Furthermore annual radar precipitation sums displayed some measurement errors. In Figure 23 and Figure 60 (Gaula and Surna) the different precipitation amounts measured at the edge of two radar beams can be found. This may be caused by partial beam blockage, attenuation and different beam heights at the border. The annual radar precipitation of Nausta shows partial beam blockage (see Figure 57). The white spots in all precipitation sums could be caused by clutter correction.

The same pattern was found when looking at the annual precipitation sum of KliNoGridRad (Figure 24). Although this data includes gauge information, beam blockages, attenuation and clutter corrections are still visible. KliNoGridRad only combines radar and gauge data, but does not consider other mistakes such as beam blockage. That means that the product leads to an improvement of the precipitation amount but cannot improve errors like shielding or clutters. Because of the discussed artifacts, radar data was not used as an input for a hydrological model, because these errors would bias the modelling results. This was also discussed and seen by Abdella et al. (2012).

Looking at the correlation between NWP and seNorge2, as well as the correlation with the weather station revealed that temperature has in general a high correlation. This high correlation can be explained, because the spatial variability of temperature is in general smaller than precipitation (Ledesma and Futter 2017). An explanation for that are large scale atmospheric temperature anomalies, that are especially spatially representative in flat areas (Orlowsky and Seneviratne 2014). When dividing the datasets by temperature and looking at the sub-catchment scale, the correlation of temperature decreases. It needs to be considered that only NWP data was used to divide them, leading to a bias that might explain this decrease. Temperatures below -5 °C seem to have the lowest correlation which leads to the assumption that the interpolation of temperature data is more erroneous with low temperatures. This might also explain the decrease of temperature correlation in higher elevated catchments because the mean temperature is decreasing with the elevation. The catchments with the lowest temperature correlation are Usma and Nausta. A reason for that might be the use of different gauging stations during the interpolation or different assumptions about the elevation influence. Compared to the weather station, lower temperatures appeared to be slightly underestimated by seNorge2. There is a spread between weather station and datasets of about 10 °C. This is an uncertainty that might affect snow melt at around 0 °C. Nevertheless, it can be assumed that temperature data is a good estimation of the ground truth for both of the datasets, because of the high correlation. Additionally, it proves that both of these datasets can be used as input for hydrological models without having a high influence on the modelling results.
Hourly precipitation has a much lower correlation than hourly temperature. This is first of all caused by the higher temporal and spatial variability of precipitation. When looking at daily precipitation, this effect decreases, because there is less spatial variability (Lussana et al. 2016a). The correlation decreases for snow. A reason for that might be the influence of wind-induced under catch that is corrected in NWP, but not in seNorge2. Another reason might be attenuation of radar beams (NWP), measurement errors caused by delayed snow melt or snow drift in personal (NWP) or public (both) gauges. Compared to the weather station, hourly precipitation did not have a good agreement even so NWP is based on radar measurements that should have a better agreement with the hourly spatial variability of precipitation. This hints to the assumption, that the radar and numerical model information included in NWP is still not enough information to estimate hourly weather processes in a mountainous region (where the weather station is located). Interpolating gauge data is also not suitable for hourly estimates, because the spatial variability is too high compared to the scarce measurement network. It must be mentioned that these results cannot be applied for a general statement of the quality of the datasets, because it is only a case study at one specific point and not necessarily true in other locations. Still, it might give a hint about the quality of the models.

At last, the precipitation amount estimated by seNorge2 and NWP was compared with inverse simulated precipitation data from discharge data. The results show that the inverse model always estimates the highest amount of precipitation. Although this might be the case, it is more likely that

there is an error in the precipitation simulation. A reason for that might be the small discharge fluctuations (caused by measurement errors) that get interpretated as precipitation from the inverse hydrological model. That way, small precipitation amounts are estimated over a long time which causes an increase of the precipitation sum. The only basin with a comparable precipitation amount between inverse model and NWP is in Gaulfoss. The reason for that might be that it is a large catchment and high discharge amounts, causing less fluctuations in the measured values. Because of the high overestimation of precipitation in some catchments, this evaluation method is not conclusive. Nevertheless, using a hydrological model was again applied for analysing individual events. It was assumed that there, the response of the basin is caused by precipitation and not by measurement errors or random fluctuations of discharge measurements.

All in all, NWP precipitation is closest to the one assumed by looking at the water balance, seNorge2 and radar data are underestimating it. Temperature data showed slightly better correlation with NWP but is in general in good agreement of both datasets and the weather station. Nevertheless, the spread of values has to be considered, especially for snow melt calculations. The correlation of the datasets decreases with higher elevations for most of the catchments. Hourly precipitation shows poor similarity to a ground observation near Usma. Daily precipitation has a better agreement. An inverse modelling approach did not lead to conclusive results when applied in a long time period.

6.2 Conclusions from calibration experiments

The calibration process is important for this study because both datasets and each catchment were calibrated individually. If the calibration result is to be considered variable for each run, the results and especially the modelling performance cannot be compared with each other, because they could depend on the calibration and its randomness.

In a first step, two calibration methods were tested and compared. DEMC achieved better model performances in terms of KGE in calibration and validation with a lower amount of simulations compared to the progressive Monte Carlo method. Furthermore, 9,000 simulations led to more stable results, especially in terms of validation. However, the method creates output files for each simulation that need a lot of computer storage. Additionally, the wall time of this method was higher than for progressive Monte Carlo that led to a lower model performance. The KGE increases with the number of simulations. Only calibrations with more simulations created stable performance results when repeating the simulation. This was also found out by Saldaña Espinoza (2022).

Because of the better efficiency and performances achieved with DEMC, this method should have been chosen. Unfortunately, there was an error with the calibration parameter output. Here, the optimal parameter set was not written into `respar.txt`, but only into `bestsims.txt`. That's why the validation of this method originally led to low validation performances. It was assumed that because of the strong variance between each simulation step, the resulting parameter set was only suitable for that specific time period instead of the whole period (including the validation time). Progressive Monte Carlo wrote the best parameter set into `respar.txt` and was therefore applied correctly in the validation and originally resulted in a better validation performance. That's why this method was applied in this study. The best results in terms of KGE and stability were achieved with 25,000 simulations per calibration. This turned out to be acceptable because testing DEMC at a later point showed that it did not improve the model performance compared to the stepwise calibration approach applied in this study.

Further analysis of calibration improvements led to a stepwise approach. Here, parameters were fine-tuned after an initial calibration and the stability of the results were tested. In the first tests of these methods and a smaller number of simulations in the initial calibration, high increases of KGE were achieved. Because there was a chance of biasing the model towards one specific process, the number of simulations was increased to 25,000 for the initial step. That way, already stable results are created that can be fine-tuned afterwards. Increasing the stability of the initial calibration led to a smaller improvement of KGE during the stepwise calibration. This confirms the law of diminishing, that says that the amount of work put into a production, will first increase up to a certain point, but after that the result will decrease (Shephard and Färe 1974). In the case of a calibration, the results are not decreasing, since the best simulation is saved. But still it can be compared that after a certain amount of work, the results cannot be increased, because an optimum performance is found.

For multi-gauge catchments, using the median KGE of all gauges led to a lower performance in terms of KGE in general and especially in Gaua (KGE of 0.292). When calibrating the model with the median, the objective function can be minimized by creating suitable results only in four of the catchments, instead of all of them. This happened in Gaula, because three stations were not fitting as well as the other four. The median is not a good calibration strategy when interested in a good agreement in every of the sub-catchments, because it does not integrate all the stations and doesn't take outliers into account. Using the average KGE improved this. Here, all gauges performed more or less equally and none of them had a KGE lower than 0.6, because the performance of all stations is included in the objective criteria. It should be mentioned that these tests were done with a small number of simulations (2,000) for each calibration. That's why an optimum was not necessarily found. Still, the general behaviour of the objective functions can be seen.
When excluding stations from the calibration and only calibrating a certain amount of them, the average KGE of all stations decreases (when using 2,000 simulations). It still must be mentioned that the decreases is acceptable, meaning that calibrating some gauges only can be a good method, when reducing computing effort plays a role. For the purpose of this master thesis (evaluating input datasets) this approach didn't seem to be reasonable, because a good suitability of every station is of interest. When only calibrating one gauge, the number of simulations was increased. That way, better performances were achieved. An interesting finding is that calibrating smaller sub-catchments and tributaries (Gaua, Lillebudal, Hugdal bru and Killingdal) resulted in a better performance for all stations than when calibrating downstream stations that get a lot of inflow from other catchments (Gaulfoss, Gaula ovf. Fora). A reason for that might be that calibrating the output of smaller sub-catchments is a better approach than calibrating the outflow of the whole catchment and applying these parameters for smaller sub-catchments. This leads to the assumption that it is better to go from a small scale to a wide scale than the other way around. Figure 31 shows a good example of that behaviour. When calibrating Gaua, a good performance of every station is achieved (especially Gaulfoss, which is the station most downstream). When however calibrating Gaulfoss, only stations directly in the Gaula river showed reasonable results, where as in contrast the tributaries perform rather poorly. Note that these assumptions are based on the KGE distribution and relative values. The absolute values should not be interpreted, because the number of simulations differ between Gaua and Gaulfoss.

Finally, the connection between low flow conditions and the choice of input data was evaluated. It turned out that NWP has a worse agreement towards low flows than seNorge2. One possibility for that might be caused by a timing issue, when the precipitation data leads to simulated discharge peaks in observational low flow conditions. Another reason might be that seNorge2 was used to calibrate the soil thickness and therefore has a better agreement of the soil recession. It was

assumed that the poor performance of low flow simulation is caused by wrong assumptions about the boundary conditions. However, this could only be proved partly, by applying GLUE or changing them manually, by a small increase in the performance compared to the initial calibration. However, the parameters seemed not to converge against specific values during the calibration.

6.3 Calibration results

By calibrating the number of soil layers and their thickness (chapter 5.2.2) catchment characteristics could be taken into account. It can be assumed that for example higher elevated catchments with steep slopes have less or thinner soil layers, because they do not have slow recession curves and do not store a lot of water except for snow and lakes. When looking at the results in Table 12, this behaviour was found in Grunnåi. Although Rinna had the best performance with three soil layers, the performance was also agreeable with only one layer. This makes sense because Rinna (as a part of the Surna catchment) is also highly elevated. Every other catchment showed the best performance with three soil layers. One reason might be that three-soil layer increase the number of parameters (the ones that had to be added to simulate the third layer) and therefore improving the fit of the model. Another reason might be that a third layer increases the possibility to hold water back and increase the recession of the catchment.

It must be considered that the chosen calibration method is not stable within each calibration (see Table 10). This means that results that do not vary strongly cannot safely be defined as the best result. Still, some KGEs vary more strongly with different soil layers, for example in Nausta. There, a dependence on soil characteristics and model performance can be assumed. Surna and Grunnåi do not vary that much in their performance. This might be caused by a lower sensitivity of these catchments towards soil characteristics.

In the next step, the stream-depth was calibrated as well by varying it by 25 cm above and underneath the ground. The stream-depth divides groundwater into standing water and water that participates in the flow of the catchment. Every catchment showed the best performance when the stream-depth was similar to the soil ground, meaning there is no standing water inside the catchment. This could be physically explained by the lack of standing water in Norway but doesn't make sense considering the high number of marsh-land-areas in the catchments. Another reason might be caused by the input data. When the precipitation amount is underestimated, the model automatically tends to include every amount of water into the discharge and doesn't participate in evapotranspiration. This theory was tested by changing the input data product in Nausta to NWP and repeating the stream-depth calibration. In that case, the best performance was achieved with a stream-depth 25 cm above the ground. That's why there seems to be a dependence between the precipitation amount and the best stream-depth performance. A possible reason for that might be that the precipitation quantity is higher and more water can be stored inside the catchment and evaporate.

6.4 Model results

In this section, the model results of each catchment and with different input datasets will be discussed. Using weather station data directly showed acceptable results in Usma, but had a low performance in Grunnåi. This low performance is caused by an underestimation of precipitation, especially in winter (see also chapter 6.5) probably caused by wind-induced under catch.

Using seNorge2 as input data resulted in the best model performance during the calibration, except for Usma and Gaula. However, the performance dropped when calculating the KGE of the validation period. This decrease of KGE might be caused by an overfitting of HYPE towards erroneous data, that does not define the actual catchment and results in a lower KGE when applied to another time.

Analysing cumulative discharge volumes revealed that seNorge2 almost always underestimates the observed discharge volume, probably caused by an underestimation of precipitation in general, which was already detected when calculating the water balance equation of the catchments.
NWP resulted in better performances during the validation period, even so there is a lower KGE than seNorge2 in the calibration, except for Usma and Gaula. Cumulative discharge sums revealed that the model calculated with NWP data is estimating to much discharge in Rinna and Nausta, but is in good agreement with observed discharge in Usma and Grunnåi.
In a further step, correction factors for snow (*pcusnow*) and rain (*pcurain*) were implemented in HYPE and added to the calibration. These factors add or decrease precipitation to the observations using a constant factor. Adding a correction to seNorge2 led to an improved of model performances except for Grunnåi. In Rinna, the correction of seNorge2 improved the KGE by 0.15, in Usma by 0.1. The performance of the validation was increasing as well. This means that correcting the input data creates more security towards other periods. NWP correction led to worse agreement of the cumulative volume sum compared to observations. It can be assumed that NWP in general is in better agreement with the actual quantity of the precipitation and doesn't need a correction, especially a constant factor. It is even possible that correcting rain might cause errors, because an over- and underestimation can vary from event to event whereas the correction factor is applied constantly for all of them. Another interesting finding is that a precipitation correction does not necessarily increase the model performance. The reason for that is probably that the whole catchment is corrected by a constant factor, although the precipitation amount is not always over- or underestimating the discharge in a similar matter.
Generally, it should be discussed whether precipitation correction with constant factors are an agreeable method to use in a hydrological model. As discussed in chapter 2.2, many measurement errors can cause uncertainty in the precipitation estimation. These errors are however often not constant over time and space. For example for precipitation gauges, wind speed plays an important factor for under catch. This process is not constant, because wind speed is variable over space and time. Another uncertainty is the spatial variability of precipitation, which can also not be accounted for with a constant factor, which was seen in Gaula, where adding a constant factor to the whole catchment did result in a poor agreements with the water balance in some of the sub-catchments. Additionally, the boundaries during the calibration of these factors should be discussed in more detail. This study assumed that precipitation can be decreased or increased by up to 50 %. However, Wolff et al. (2015) stated that snow can be underestimated by up to 80 % for unshielded gauges. Another question might be whether precipitation can actually be overestimated by the datasets. In terms of radar data, an overestimation might be caused by bright bands, a wrong Z-R relation or attenuation correction factors. Gauge data can be overestimated by systematic errors. Whether these errors should have the same range as underestimation needs to be further discussed.

Some model results did not seem to be strongly influenced by the choice of input data, but are anyway worth mentioning here. The first finding was a decrease of model performance in the winter month in almost all catchments. A reason for that might be the snow melt model of HYPE. There are different models available, but most of them are meant for a daily timestep. Only snow scaling can be applied to the hourly model instead of the default. This model option was tested for one catchment but did not result in a much better performance in winter. That's why it was not applied in this study.
Another reason, besides a model uncertainty, might be that temperature data has a higher influence in winter than in summer, because temperature is the main factor of snow melt (SMHI 2021). An error in the temperature data might be the reason for a poor performance in winter. Temperature data is known to be more accurate because of the lower spatial variability. Nevertheless, a bias in the temperature data around the melting temperature might be a source of

error. Another reason of poor model performances in winter might be caused by measurement errors in winter, for example precipitation measurements tend to be less accurate for snow, because of wind influences near gauges or wrong estimated Z-R relations of radar data. That might cause erroneous precipitation data that is not able to simulate the snow melt after it accumulated over winter. Another measurement error could be in the discharge data. In Norway, river tend to build ice during winter. That might cause errors when measuring the water level. Another uncertainty might be that low flows are dominant during winter which are usually underestimated by the model. This will also be discussed in this chapter.
Nausta however did not loss its model performance during winter. This can be explained by the fact that Nausta is located in the Atlantic regime (Gottschalk et al. 1979) and is more influenced by precipitation and evapotranspiration rather than snow melt.

Another finding was the underestimation of low flows in most catchments, especially in Nausta, Grunnåi, Gaula and Surna. In these catchments, none of the datasets show a better agreement compared to the other. That's why it can be assumed that the low flows do not mainly depend on the input data, but are influenced by another source within the model. A reason for the underestimation was however not found in this study.

It can be concluded that seNorge2 often achieves a better model performance than NWP, however, the validation performance is usually lower. The model performance improved even more when using a correction factor for seNorge2. This leads to the assumption, that the data is not wrong in general, but has a measurement bias (for example wind-induced under catch) leading to an underestimation of precipitation. However, adding a constant factor to the measurement is a strong simplification. The decrease of KGE in the validation period might be caused by an overfitting of the model, meaning that the input data is not correct, but the model is just using unrealistic estimates, that do not work for another time period. Furthermore, the precipitation correction factors might not be time independent. In that matter, stable NWP results seem to be more reliable. Using different datasets as input led to a wide spread of simulated discharge, especially in small scale catchments like Usma and Grunnåi. This is in agreement with Bárdossy et al. (2022), where precipitation data was considered to have a high influence on the uncertainty of a hydrological model.

6.5 Water balance

Looking at the water balances of the modelled results helped to evaluate how HYPE handles the different input datasets. It was found that HYPE compensates precipitation underestimation by estimating a low amount of evapotranspiration to avoid water loses out of the catchment. This happened in every catchment when using seNorge2 as input data. This could also be explained physically, because there is no actual evapotranspiration when no water is available in the system. However, the catchment are known to have marsh areas and should therefore not be dry during summer.
Because the lack of water is compensated by the evapotranspiration, seNorge2 simulates a similar amount of discharge compared to the measurements in Nausta and Grunnåi. In Usma and Surna, even decreasing the evapotranspiration is not enough to match the observed discharge volume. The corrected seNorge2 version improves the water balance by adding more water into the catchment. With that, the model results are closer to the observed values in terms of volume. Even the evapotranspiration gets closer to MOD16 values, except for Grunnåi, where the evapotranspiration doesn't change. SeNorge2 does not lead to a storage change within the model, but when calculating the storage change with measured discharge, there is a strong decrease of

storage in Usma and Surna (Rinna). When looking at the seasonal changes, it becomes clear that seNorge2 always underestimates the discharge, especially in winter. Except for Grunnåi, where seNorge2 is higher in August and September. Correcting seNorge2 does not lead to a better suitability, because the differences vary between the months (sometimes over-, sometimes underestimating), except for Nausta, where the measured discharge amount is in good agreement with the model using a correction factor with seNorge2.

NWP is overestimating the discharge in Nausta and Surna. The evapotranspiration amount is close to MOD16. This shows that when enough water is available in the catchment, the model calculates more evapotranspiration. A complete overestimation of AET is not possible, because it is bound by the PET calculated with Hargreaves. This may also explain the water storage increase in Usma and Nausta within the model. Here, the precipitation amount could not be compensated by evapotranspiration and discharge and led to an increase of water storage within the modelled catchment. This storage changes were described in chapter 5.3.3.3. In Usma, the water accumulates through the river, probably caused by a high dampening effect leading to unrealistic water accumulations in the main river. A reason why these accumulations might be possible is because of the relatively small time period where these storage changes can already be seen but will not have as much influence as in a later timestep. In Nausta, water is accumulated by snow in the mountains, that is not fully melting during the year. A reason why that happens in Nausta might be that NWP estimates more precipitation in the upper catchment compared to the lower basin (see Figure 59). Maybe the high amount of precipitation in the higher elevated areas led to the accumulation of snow in the upper catchment. When looking at seasonal changes, the model with NWP appears to also have a good seasonal agreement with MOD16. In most catchments, NWP leads to overestimating the discharge amount of the spring melt which might be caused by an overestimation of snow during winter. E.g., in Surna, the difference between measured and simulated discharge is variable between the months (e.g. overestimated in August and underestimated in September).

At last, weather station data was included in two catchments (Usma and Grunnåi). They both seemed to underestimate the precipitation amount and led to an underestimation of discharge volume. This was especially high in Grunnåi. The precipitation amount is especially low in winter, probably caused by wind-induced under catch of snow. Another reason might be that weather station data is not spatially available and changes of precipitation because of elevation were not considered.

The evaluation of the water balance in Gaula showed that seNorge2 is underestimating the precipitation amount in all catchments which is compensated by HYPE through low evapotranspiration. NWP is sometimes over- and underestimating the precipitation. When precipitation is overestimated, some models started accumulating water in the catchment's storages. Correction factors improved the annual water balance, but showed under- and overestimations when looking at seasonal changes.

6.5.1 Comparison evapotranspiration MOD16 and Hargreaves

This section is going to analyse the evapotranspiration calculated from Hargreaves. Hargreaves only needs the minimum and maximum daily temperature as well as the latitude for the calculation. By using seNorge2018, an independent dataset was applied with data available over the whole time period. Using this extra data source might be an unnecessary extra step but was found to be very useful because evapotranspiration is an important part of the water balance. Especially considering

that this study found out that HYPE tries to avoid water accumulation within a catchment by adapting the evapotranspiration amount. That's why the model would calibrate the evapotranspiration depending heavily on precipitation (depending on whether it over- or underestimates the water amount). By creating boundaries with PET, an overestimation of evapotranspiration was avoided. Calculating AET could also improve that, but it depends on the water availability within the catchment and is more meaningful when calculated within the model. Comparing MOD16 PET with Hargreaves, showed that Hargreaves always estimated lower amounts. This does however not mean that the calculated PET is underestimated, because it is known that MOD16 tends to overestimate PET in Norway (Huang et al. 2019). Comparing PET from Hargreaves with AET from MOD16 showed that the estimated PET is still higher than AET estimated by MOD16. That is a good sign, because it can be interpretated as if HYPE has the possibility to simulate the same amount of AET than measured by MOD16 when enough water is available and that the PET boundary is not underestimating the AET from MOD16.
From the previous result, we already saw how important the evapotranspiration in the model is, because it is used to balance the available water amount in the catchment. If this is not possible, the storage volume increases. It was an important step to apply PET as an additional model input, because otherwise (when the PET parameters would have been calibrated) the model would have probably reacted with high evapotranspiration amounts when precipitation is too high. This is what it did with underestimated precipitation as well.

6.6 Events

After looking at the models in general, selected events were further analysed. At first, three events where the discharge peak was not simulated were shown. NWP was missing precipitation data in one event, seNorge2 at two other events. To check whether the model was not reacting appropriately to the discharge, an inverse modelling approach was applied that estimated the actual precipitation amount using discharge data. In some cases, this showed that the precipitation was comparable for both datasets, but the model response wasn't (Figure 94), in other cases precipitation was missed or underestimated (Figure 40). That seNorge2 especially missed convective events seems reasonable, because it is only based on point data, that is more likely to miss convective, small scale precipitation cells. This was the case in Usma in July 2015 (Figure 40). NWP, that includes radar data, measured this convective cell. Nevertheless, this did not lead to a better model fit, which might be caused by a model uncertainty. Another reason might be that the precipitation amount from NWP wasn't high enough because there was more precipitation volume assumed by the inverse modelling approach. This event is especially interesting, because even the weather station, located 30 km away from the catchment, did not observe the precipitation. That radar can have an advantage compared to gauges for small scale events was also shown in Reinemo (2012).

Another reason for missed discharge peaks in the models is temperature and snow melt. Two events with that behaviour were shown (Figure 95 and Figure 96). Here, the reason for a mismatch with observations could be erroneous temperature data, leading to a lack of snow melt, that was actually higher than assumed. Another reason might be that snow melt also depends on radiation and albedo effects (at least for old snow, that is dirty and has therefore a lower albedo), which is not considered by the model and could lead to a snow melt within the catchment even with temperatures below 0 °C. Another possibility might be that the discharge increases might be caused by unnatural discharge changes, e.g. through hydro power plants. This should however not be the case because the catchments are considered to be unregulated.

The last finding are events with discharge peaks calculated by the models that were not measured in reality (Figure 97). In one of the events seNorge2 estimates a high precipitation amount, although there was barely any reaction from the catchment. This might be caused by a precipitation cell over a nearby located weather station, that was included in the interpolation of seNorge2, and led to a wrong estimation on what happened inside the catchment. Another reason might be that the model used with seNorge2 is more likely to react stronger with precipitation, because this product is usually underestimating the water amount. In another event, NWP estimated precipitation leading to a wrongly simulated discharge peak, that wasn't simulated using seNorge2 (Figure 98).

To summarize this chapter, it can be said that differences between simulated and observed discharge can have various sources. Here, events were shown where simulated discharge differed from observed discharge because of wrong precipitation estimation (over-, underestimation or completely unmeasured precipitation cells), temperature or modelling errors.

6.7 Sensitivity analysis

The last part of this analysis was an evaluation of the sensitivity of the models for both datasets and the parameters. GLUE was applied to test the equifinality and to find sensitive parameters. The results showed a wide spread of some parameters visualizing the equifinality of the model. Some parameters were more sensitive than others. For example, the parameter *rivvel* (celerity of flood in watercourse) had similar values in each calibration and not a wide spread. This means that this parameter is very sensitive and always lead to a convergence on one specific value. Other examples of similar behaviour are *pcurain* (under catch correction for rainfall), *pcusnow* (under catch correction for snowfall) and *damp* (fraction of delay in the watercourse which also causes damping). These values are less sensitive in Nausta, which might be caused by a less sensitive catchment in general, maybe caused by the catchment size.
Other parameters were found, that did have variable values for each run, but had only a small spread of values within the best five results of each run. These parameters can be interpreted as sensitive within the calibration, but are influenced by a combination of parameters. An example for that are the lake parameters *grata*, *gratp* and *gratk* (parameters of rating curve for lake outflow). All of them are applied to describe the lakes storage-discharge relation. There are different combinations possible, that all describe an equal relationship and are more sensitive within the combination than in only one specific value. Other parameters with that behaviour are *cmelt* (melting parameter for snow) and *ttpd* (threshold temperature for snow-/rainfall), which can also be seen as a combination, because the snow melt is depending on the temperature threshold and a snow melt factor.
The last category are parameters that did not even converge within a calibration yet alone with each run. These parameters must be very unsensitive because they do not converge at all. An example of that is *cevp* (evapotranspiration constant). This parameter would probably have a higher influence if PET wouldn't already be included into the model as input data. Other parameters like that might also just have a small influence on the output or maybe the number of simulations was not high enough for these parameters to convergence. This unsensitivity can mostly be seen on SLC dependent parameters. A reason for that might be that the catchments are too small to create a reasonable assumption of land use dependent parameters. A combination of catchments might create more reasonable results.
To test whether the parameters converge against the defined boundary conditions, the parameters where normalized and visualized as well. It turned out that there were no parameters that converged to the boundary conditions, except for *pcusnow* and *rivvel*. Because *rivvel* was converging against zero, a change of boundaries is not necessary. SeNorge2 calibrations often

achieved *pcusnow* values close to 0.5 (maximum boundary). This shows how strongly snow is underestimated by this product. Furthermore, the parameter space of the two datasets was compared. In Grunnåi, most parameters were assumed in an equal range by both datasets. Differences are especially found in *pcusnow* and *pcurain*, which seems reasonable considering that they are dependent on precipitation data. Another parameter is *deepmem*, which describes the memory of the soil. SeNorge2 converges to higher values here. This makes sense considering that seNorge2 tends to underestimate precipitation and keeps water in the soil longer to better assume discharge peaks in the next precipitation event.

After analysing the parameter space and equifinality, the influence of the calibration period was evaluated by switching calibration and validation periods. In all catchments, there was a difference in the KGE when switching the calibration and validation period. This means that none of the models calculates results that can be uniformly applied to another time without a decrease or change in their performance. However, some results varied more strongly than others. In Nausta, a model with seNorge2 originally had a higher decrease in the validation period. When switching, the validation performance increased drastically. This might be a coincidence from the calibration. Another possibility is that the calibration of period 2 is a better representative than period 1. NWP in Nausta leads to stable results during calibration and validation, even with changed periods. It can be assumed that both datasets are assuming precipitation data correctly here.
The difference is higher in Grunnåi. SeNorge2 leads to a strong decrease of performance in the validation, also when changing the periods. It can be assumed that the model is overfitted for that catchment and creating a good performance during the calibration, that is however not appliable for other periods. NWP in Grunnåi leads to more stable results. This proves, that HYPE is able to model this catchment and it can be assumed that the main source of error in the seNorge2 model is the input data. In Rinna, the same findings were seen, except here, NWP is the input data that created to unstable results. By calculating a double sum curve, the reason behind this behaviour was found. Depending on the chosen time period, the correction factor of precipitation changes. This means the value is not constant over time. In Rinna, it seems like the correction factor is more influenced by the time with NWP than seNorge2, because the precipitation amount is changing more with that product.

In the final step, parameter results were evaluated. The rain correction factor varies in each catchment and dataset. SeNorge2 is almost always close to zero or above zero, meaning that the rain is underestimated and is increased in the model. This is also the case for snow correction. This makes sense, because as discussed in the mass balance of the catchments, seNorge2 is underestimating precipitation in every basin. That snow correction is higher than rain is reasonable considering that wind-induced under catch has a higher influence on snow than on rain. Rain and snow correction of NWP as a model input resulted in positive and negative correction factors depending on the catchment. That the under- and overestimation of precipitation with NWP is varying with each catchment was also detected in the mass balance analysis (chapter 5.1.4).
Another comparison was done with the river velocity. The results show that this parameter is varying in each catchment but does also depend on the input dataset. The maximum velocity is 3 m/s which is high and unrealistic for a catchment like Nausta. A reason for high river velocities might be that the catchment reacts fast to precipitation events and the model tries to balance that with high velocities. Because the river velocity is behaving unrealistically, it can be assumed that the river velocity is compensating another process. The reason might be the input data, because the value changes when using different datasets. That the river velocity cannot be explained physically can also be seen when the catchment slope is compared to the parameter value, because in a realistic system, these values should be correlated. However, this was not detected.

Another parameter that represents a physical process is the snow melt factor *cmelt*. This value is SLC dependent. It can be expected that the parameter represents the characteristic of these land uses. However, the parameter varies depending on the dataset and catchment. In Beldring et al. (2003), a parameter estimation for Norway was done using a HBV model. Snow melt factors for different land uses were estimated. Here, the forests had a value, depending on whether they are high or low elevated between 0.16 and 0.1 mm/°C/hour. Mountainous areas had a snowmelt factor of 0.14 mm/°C/hour. These values are in a same area as the calibrated values from HYPE (they are between 0.09 and 0.4), but do not have the same land use characteristics. The factor does also depend on the input dataset. This can be seen in the mountain SLC, where high values are assumed with seNorge2 and lower values with NWP. In other SLCs, the values are more similar, for example in the Urban/Agriculture SLC. According to Beldring et al. (2003), the snow melt factor is the highest in mountain areas as well as high elevated forests. That is in agreement in Usma and Rinna, but not for the other catchments. This makes sense, because these catchments have high elevated forest areas, whereas the other catchments mostly have forest SLCs in lower elevated areas. It can be concluded that there might be some physical truth behind the parameters, but because they were calibrated for only one catchment, they probably not always converge to the actual characteristic of the SLC.

This sensitivity analysed evaluated different model uncertainties, caused by calibration, precipitation data, choice of calibration period and possible overcompensations of parameters as well as unrealistic values. This chapter discussed their possible influence on the model results. To compare the uncertainty of the model with uncertainty of the input data, the sensitivity analysis should be further researched.

6.8 Summary

This chapter summarizes the findings when comparing seNorge2 and NWP in general and using them as input data in HYPE. Table 16 provides an overview over the results of this study.

Table 16: Comparison of results from this study with seNorge2 and NWP

	seNorge2	NWP
General	Annual precipitation amount is underestimated in each catchment	Annual precipitation amount is sometimes over- and underestimated
	Low correlation to hourly weather station data	
	Temperature has a negative bias towards the weather station near Usma	
	Higher difference to other dataset with hourly precipitation and for temperatures below 5 °C	
Model results	Higher KGE during calibration in Nausta, Rinna, Gaula and Grunnåi	Higher KGE during calibration in Usma
	Higher KGE during validation in Gaula	Higher KGE during validation in Nausta, Rinna and Grunnåi
	Low flow underestimation in most catchments	
Water Balance	Underestimation of precipitation in every catchment	Varying precipitation amount
	Low AET amount	AET comparable with MOD16 data
	Discharge volume underestimation in Usma, Surna and Gaula	No underestimation of discharge amount
	No storage accumulation in catchment	Accumulating water storage in Usma and Nausta
	Precipitation especially underestimated in winter	Precipitation overestimated in winter for example in Grunnåi
Events	Some precipitation events were not measured at all	Most events were measured, but not necessarily correctly
	Some events were wrongly measured as precipitation	Some events were overestimated
	Temperature errors or other uncertainty sources led to differences between the simulated and observed discharge	

7 Conclusions

The aim of this thesis was to evaluate different precipitation products and test them as input into HYPE. This was done by setting up a HYPE model in five catchments and calibrating them with a stepwise calibration using these datasets. The results were analysed by performance, water balance and evaluating individual events. At last, a sensitivity analysis was done for two of the catchments. Radar data was analysed initially, but because of a high degree of beam blockage and attenuation in some of the catchments, it was not included in the model set-up. This seemed reasonable, because another product (NWP) including radar data is available. It is a combination of gauges (personal and public), radar data and a numerical weather prediction model as the background of a Bayesian interpolation method. According to the water balance, NWP sometimes overestimates precipitation and sometimes underestimates it. Hourly data was not in good agreement with gauge station data and also has low agreement with seNorge2 data. When applying NWP as input to HYPE, the calibration results were often suitable, but did not outperform other datasets. Nevertheless, the results appeared to be more stable. Furthermore, the modelled evapotranspiration had the best agreement with MOD16 estimations. In some catchments, using NWP led to an accumulation of water within the model, probably to even out precipitation overestimation. SeNorge2 is based on interpolated rain gauge data. In every catchment, it underestimates the precipitation amount. Looking at the water balance showed that this underestimation was especially high in winter. The same thing was found for weather station data. A reason might be wind-induced under catch of precipitation, that is higher for snow than for rain and therefore led to an underestimation. This did however not change a good performance during the calibration, it had however a negative effect on validation.

Some precipitation events discussed in this study showed an advantage when using radar-derived data, because small scale precipitation events were found, where weather station data was not able to measure it. However, the precipitation amount was often under- or overestimated by NWP or the model didn't react to the precipitation correctly.

Having a semi-distributed hydrological model over a wide scale, as HYPE is set-up in Norway, is a good possibility to evaluate gridded precipitation datasets by simulating the discharge of a catchment and comparing water balances. However, HYPE is complex and highly parameterized and therefore model uncertainties have to be analysed and taken into account. Another challenge was calibrating the model in an accurate way to create reproduceable results when calibrating the catchments with different datasets. This was especially complicated in Gaula with more than one discharge gauge, because the calibration is supposed to be unbiased towards the location. This is not the case when for example precipitation data might be biased differently depending on the position.

When testing the sensitivity of the model with GLUE, parameters were found, that are sensitive, sensitive towards certain combinations and not sensitive at all. It is possible that a higher number of simulations is necessary to optimize these values but that would have taken a lot of computing effort. When switching calibration and validation periods, some input data results showed instabilities. This was the case for seNorge2 in Grunnåi and NWP in Rinna, meaning that both datasets seem to have created unrealistic calibration results in different catchments. The rain and snow correction factor calibrated in HYPE is different in each catchment. However, it was always above zero for seNorge2, which agrees with the results of the mass balance.

This study applied different methods to evaluate precipitation data and their behaviour in hydrological models. No dataset seemed to dominate the other, but often seemed to estimate more realistic results.

8 Outlook

This study covered different approaches of precipitation dataset evaluation, but led to new questions and research approaches that could be a follow-up of this thesis.

Looking at Norwegian radar data showed the high number of error sources. Some of these effects could be decreased by a higher density of radar station. Especially beam shielding effects, attenuation and the height of the radar beam at wider ranges could be improved by that. A new radar station is already planned which will hopefully improve the datasets, although it will be challenging to find locations in the mountainous area that are not heavily affected by shielding (Lusanna 2022, personal communication, 10th March).

Other possibilities to improve radar data should be further explored. Besides combining reflectivity with temperatures, other methods to estimate the Z-R relation should be considered, because using Marshall and Palmer (1948) for rain rate estimation is another uncertainty.

In Gaula, a good model performance was achieved when only calibrating the small sub-catchment Gaua. Although these results might just be random, further tests should be done to see whether that can be repeated. A possibility for that is in the Surna catchment, when other data sources become available.

To increase the number of hourly precipitation gauges, a disaggregation of daily data could be considered. In Müller and Haberlandt (2018) daily precipitation data is disaggregated into 5 min and 1 h timesteps. This was done by dividing timesteps into smaller timesteps and distributing the precipitation data with a certain probability in a cascade model. Hourly data showed good agreement with observations. The advantage here is that there are more daily weather stations available than hourly gauges.

For further improving HYPE, limiting the model complexity and decreasing the number of parameters is suggested. A possibility is to define uncertain parameters using expert knowledge or other research. That way the calibration might be more reproducible. Because the river velocity had some unrealistic values in some calibrations, it might be useful to limit this parameter further, for example by considering catchment characteristics. Another calibration strategy could be applied by combining the automatic calibration with the validation performance in an iterative way.

Furthermore, HYPE simulation results could be compared to another model that uses more simplistic approaches to test the advantage of the model complexity. A possible model could be the modelling framework RAVEN, where the flexibility of a model can be adjusted and specific processes portrayed individually (http://raven.uwaterloo.ca/).

During the annual HYPE conference in 2022, a new model option was introduced. Ensemble based Kalman Filters can now be included into the model. That way, uncertainties and inaccuracies of input data can be considered and spatially integrated. A catchment like Gaula showed that spatial variability of input bias is important and should be considered in further analysis.

Another possibility might be to use the approach from Bárdossy et al. (2022) and apply inverse hydrological methods to iteratively calculate the precipitation amount of a catchment and correcting already available datasets with that knowledge, or at least get an better knowledge about their uncertainty.

Appendix

A HYPE

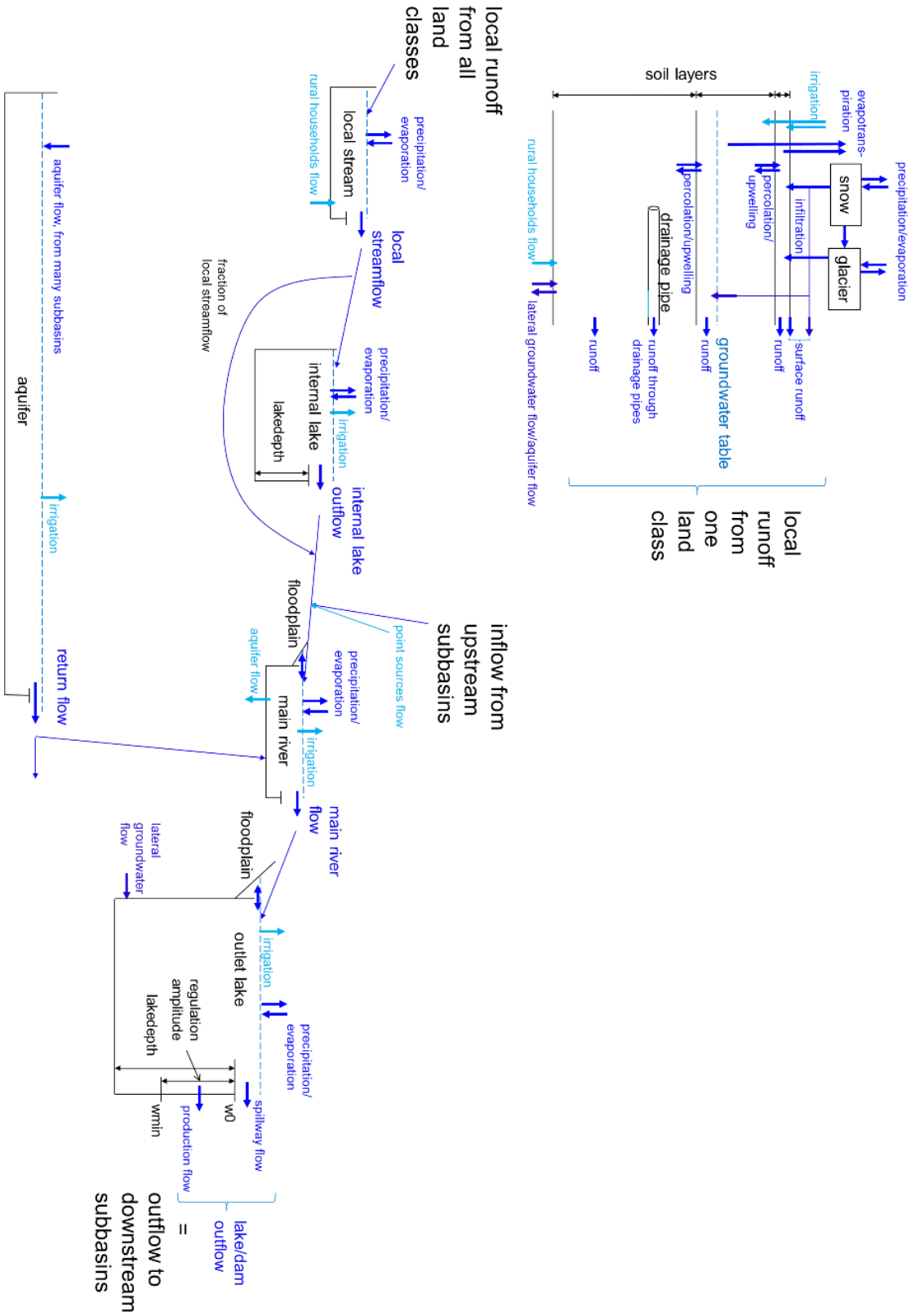


Figure 46: HYPE flow chart, figure from SMHI (2021)

A.1 Model folder structure and file examples

HYPE is a text-file based model that only consists of an executable file (`HYPE.exe`). The rest of the files are text files specified for the catchments. `Filedir.txt` defines the working dir of the model. `GeoClass.txt` contains information about the properties of the SLCs. It combines the specification of special classes (for example olakes and ilakes) as well as soil layer properties for each SLC. `GeoData.txt` defines the characteristics of each sub-basin. It contains geographic properties like mean elevation, slope and river length. The fractions of SLCs are defined there. `Info.txt` defines all the model properties. It includes the definition of start and end date, calibration settings, model settings and required output files. In `LakeData.txt` manual lake parameters can be defined (e.g. relating to volume and rating curve). `Optpar.txt` defines the boundaries of parameters during the automatic calibration, whereas `par.txt` contains the actual parameter values when the parameters are not calibrated. Observational data is saved in separate text files. `Pobs.txt` contains precipitation data, `Tobs.txt` temperature, `Qobs.txt` discharge and `Xobs.txt` can contain different observations, for example evapotranspiration, radiation or snow water equivalent. These files are shown in Figure 47.

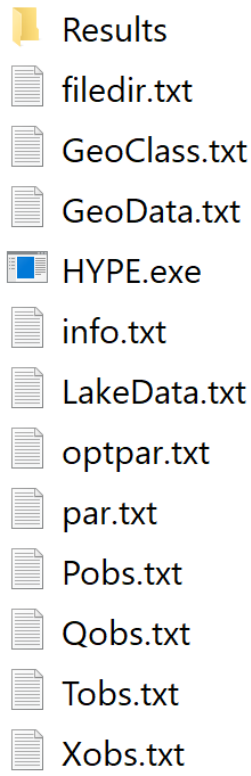


Figure 47: Screenshot of example HYPE folder

A.2 HYPE parameters

Table 17: Hype parameters description and their boundaries and step size for the calibration, descriptions from SMHI (2021)

Name	Description	Min	Max	Step size	Number of steps
ttpd	deviation from ttmp for threshold temperature for snow-/rainfall	-2	2	0.05	80
ttpi	half of temperature interval with mixed snow- and rainfall. Interval is (ttmp+ttpd) +/- ttpi.	0.05	2	0.025	78
deepmem	deep soil temperature memory	10	300	2	145
cevpam	amplitude of sinus function (about 1) that corrects potential evapotranspiration.	0.5	1.5	0.01	100
cevpph	phase of sinus function that corrects potential evapotranspiration	30	180	1	150
lp	factor for calculating the soil water limit for potential evapotranspiration	0.85	1	0.001	150
epotdist	coefficient in exponential function for potential evapotranspiration's depth dependency	0.1	10	0.1	99
sdnsnew	density of new-fallen snow (former snowdens0)	0.05	0.3	0.0025	100
snowdensdt	increase of snow density per day	0.0005	0.003	0.000025	100
fsclim	limit of fractional snow cover area for onset of snowmax	0.0001	0.001	0.00001	90
fsckexp	parameter for snowmax	0.000001	0.0002	0.000002	100
rrcs3	recession coefficient for slope dependence (upper layer)	0.0001	0.05	0.0005	100
wcfcN*	fraction of soil available for evapotranspiration but not for runoff, for Nth soil layer	0.05	1	0.01	95
wcwpN*	wilting point as a fraction, for Nth soil layer	0.01	0.5	0.005	98
wcepN*	effective porosity as a fraction, for Nth soil layer	0.001	0.5	0.005	98
mperc1*	maximum percolation capacity from soil layer 1 to soil layer 2	5	150	1	145
sfrost*	frost depth parameter	0.1	1	0.01	90
rrcs1*	recession coefficient for uppermost soil layer	0.01	1	0.01	99
rrcs2*	recession coefficient for lowest soil layer	0.01	1	0.01	99
srrate*	fraction for surface runoff	0.01	0.5	0.005	98
macrate*	fraction for macro-pore flow	0.01	0.5	0.005	98
mactrinf*	threshold for macro-pore flow	1	150	1	149

mactrsm*	threshold soil water for macro-pore flow and surface runoff (fraction of wcwp+wcfc in uppermost layer)	0.01	1	0.01	99
cmlt*	melting parameter for snow	0	0.5	0.005	100
ttmp*	threshold temperature for snow melt, snow density and evapotranspiration	0	2	0.02	100
cevp*	evapotranspiration parameter	0	0.1	0.001	100
frost*	frost depth parameter	0.5	1	0.005	100
srrcs*	recession coefficient for surface runoff (fraction), set to 1 for lakes	0.01	1	0.01	99
surfmen*	upper soil layer soil temperature memory	1	150	1	149
ttrig*	temperature threshold for soil temperature control on soil evapotranspiration	-1	3	0.05	80
treda*	soil temperature control on soil evapotranspiration	0	1	0.01	100
tredb*	soil temperature control on soil evapotranspiration	0.5	2	0.015	100
fscmax	maximum fractional snow cover area	0.5	1	0.005	100
fsck1	parameter for snowmax	0.05	0.5	0.005	90
fsceff	efficiency of snow cover to influence snow melt and snow evaporation	0.01	0.9	0.01	89
fscdistmax*	maximum snow distribution factor	0. 5	1	0.005	100
fscdist0*	minimum snow distribution factor	0.4	0.8	0.005	80
fscdist1*	std coefficient for snow distribution factor	0.0001	0.005	0.00005	98
depthrel*	depth relation for soil temperature memory	0.01	5	0.05	100
rivvel	celerity of flood in watercourse	0.01	3	0.03	100
damp	fraction of delay in the watercourse which also causes damping	0.01	1	0.01	99
deadl	parameter to calculate the dead volume in the local watercourse	0	10	0.1	100
deadm	parameter to calculate the dead volume in the main watercourse	0	10	0.1	100
gratk	parameter of rating curve for lake outflow	10	150	1	140
gratp	parameter of rating curve for lake outflow	0.5	4	0.05	70
grata	parameter of rating curve for lake outflow	0.001	10	0.1	100
gldepi	depth for all ilakes	1	100	1	99
gicatch	fraction of local runoff that goes through the local lake (ilake)	0.01	0.5	0.005	98
pcurain	under catch correction for rainfall, rainfall = rainfall*(1+pcurain).	-0.5	0.5	0.005	200
pcusnow	under catch correction for snowfall, snowfall = snowfall*(1+pcusnow).	-0.5	0.5	0.005	200

*Parameter is calibrated for each SLC separately

B Additional information about the material

B.1 Study areas

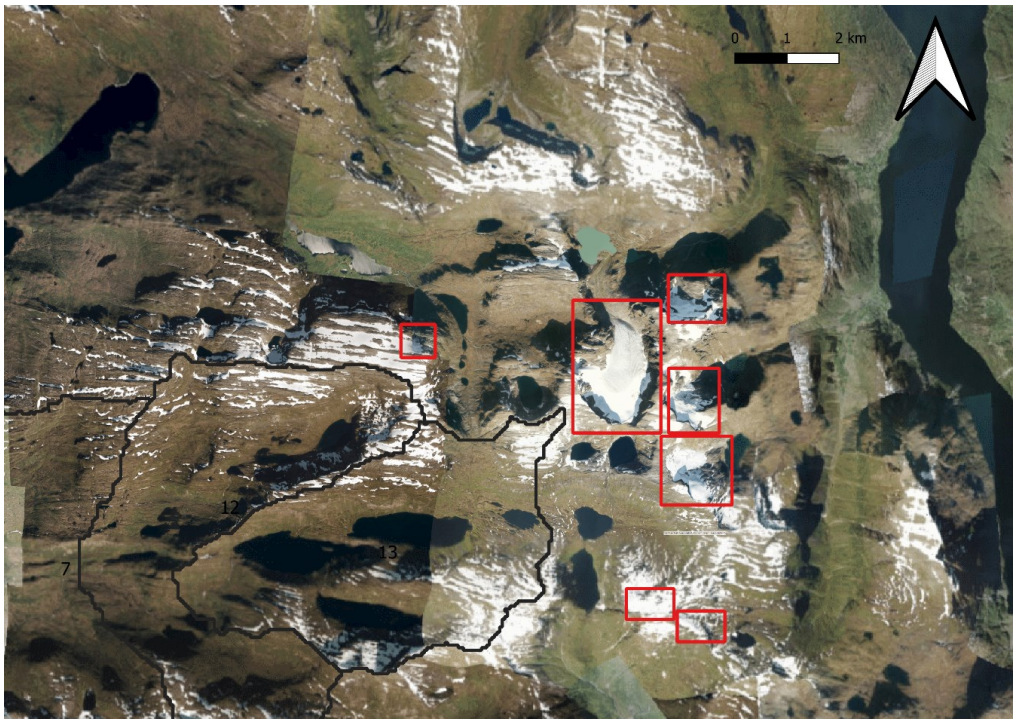


Figure 48: Locations of glaciers near the Nausta catchment (NVE Atlas)

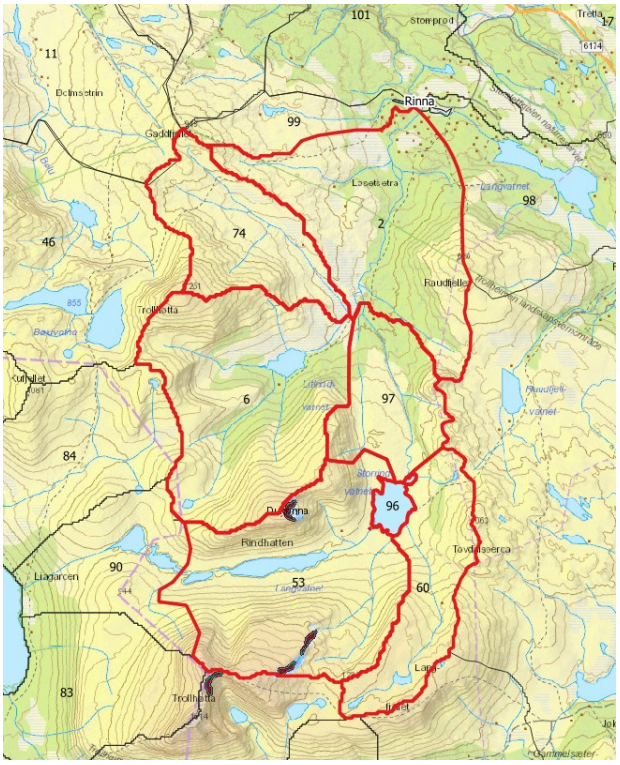


Figure 49: Location of three glaciers in the Rinna catchment (NVE Atlas)

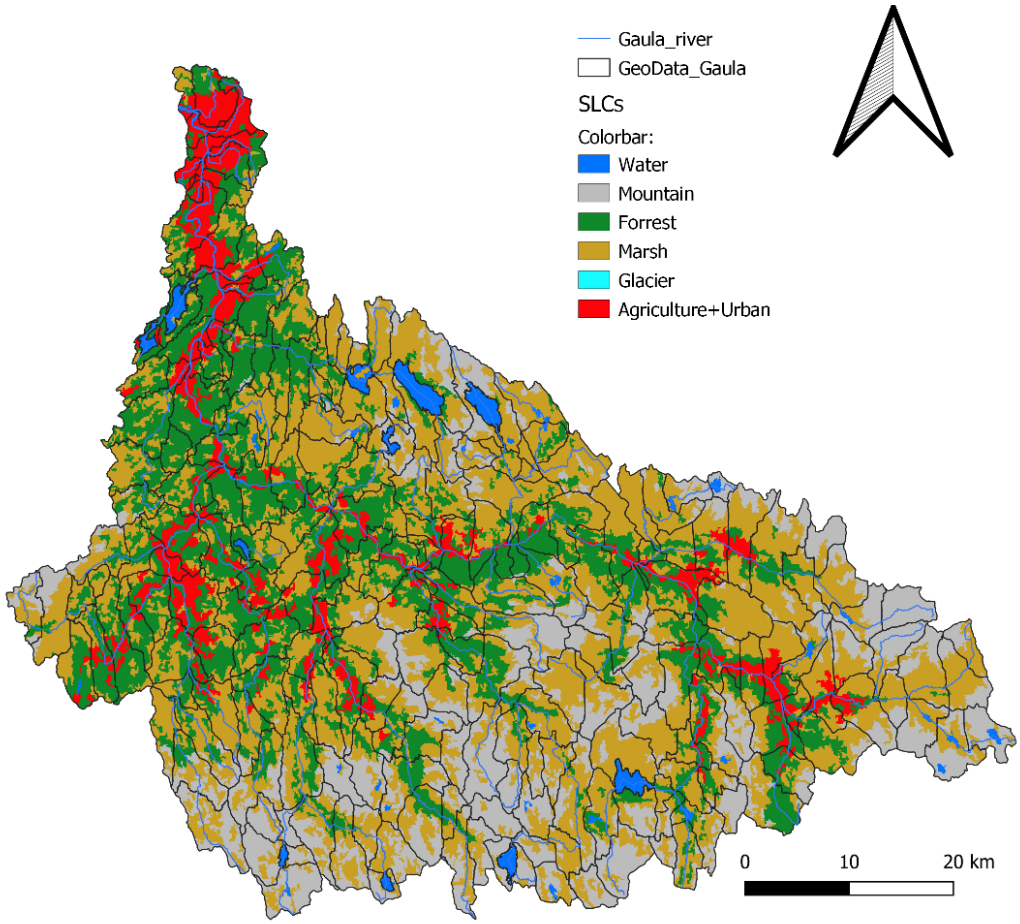


Figure 50: SLCs of the Gaula catchment

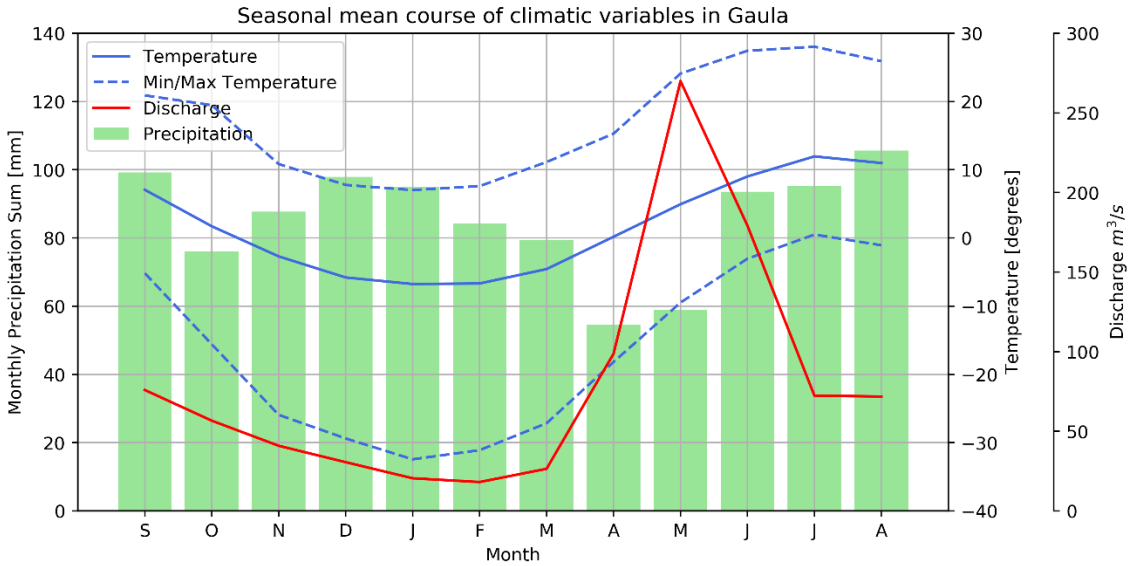


Figure 51: Seasonal mean course of the temperature, precipitation and discharge (measured in Gaulfoss) in a time span of 20 years (2000-2020) in Gaula. Data: seNorge2018 daily

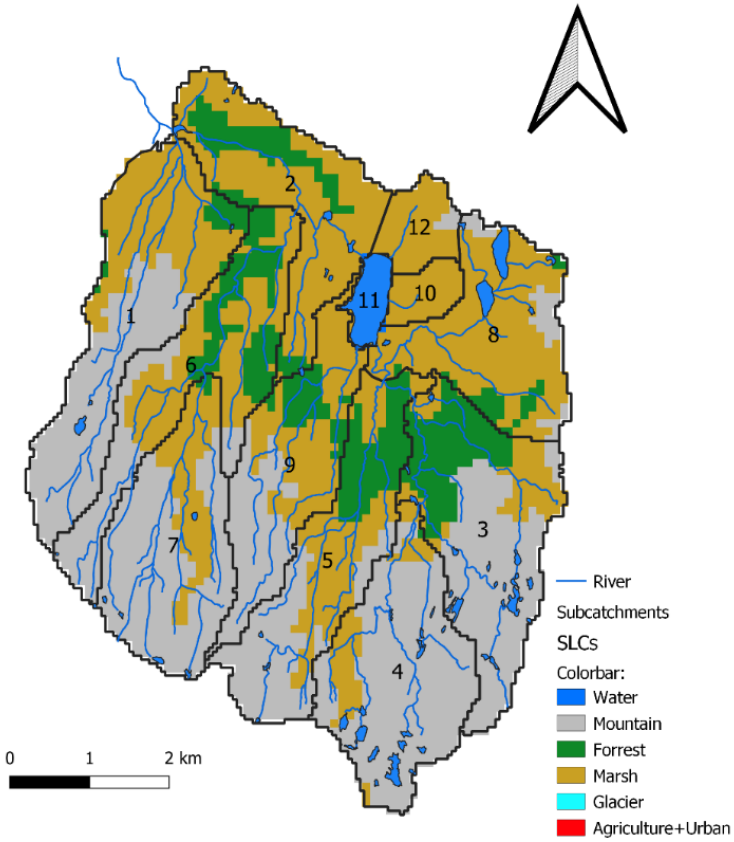


Figure 52: SLCs of the Usma catchment

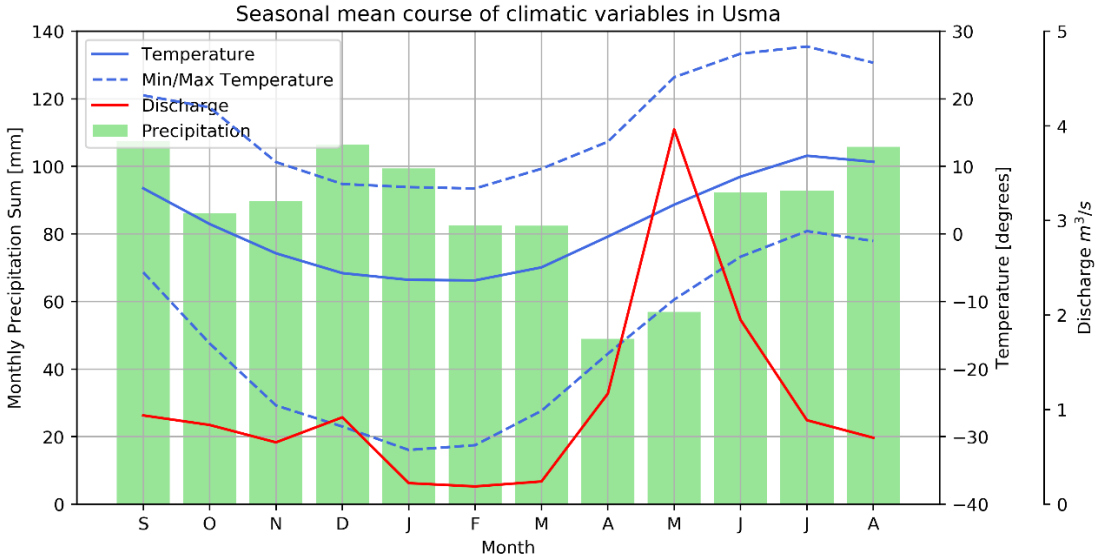


Figure 53: Seasonal mean course of the temperature, precipitation and discharge in a time span of 20 years (2000-2020) in Usma. Data: seNorge2018 daily

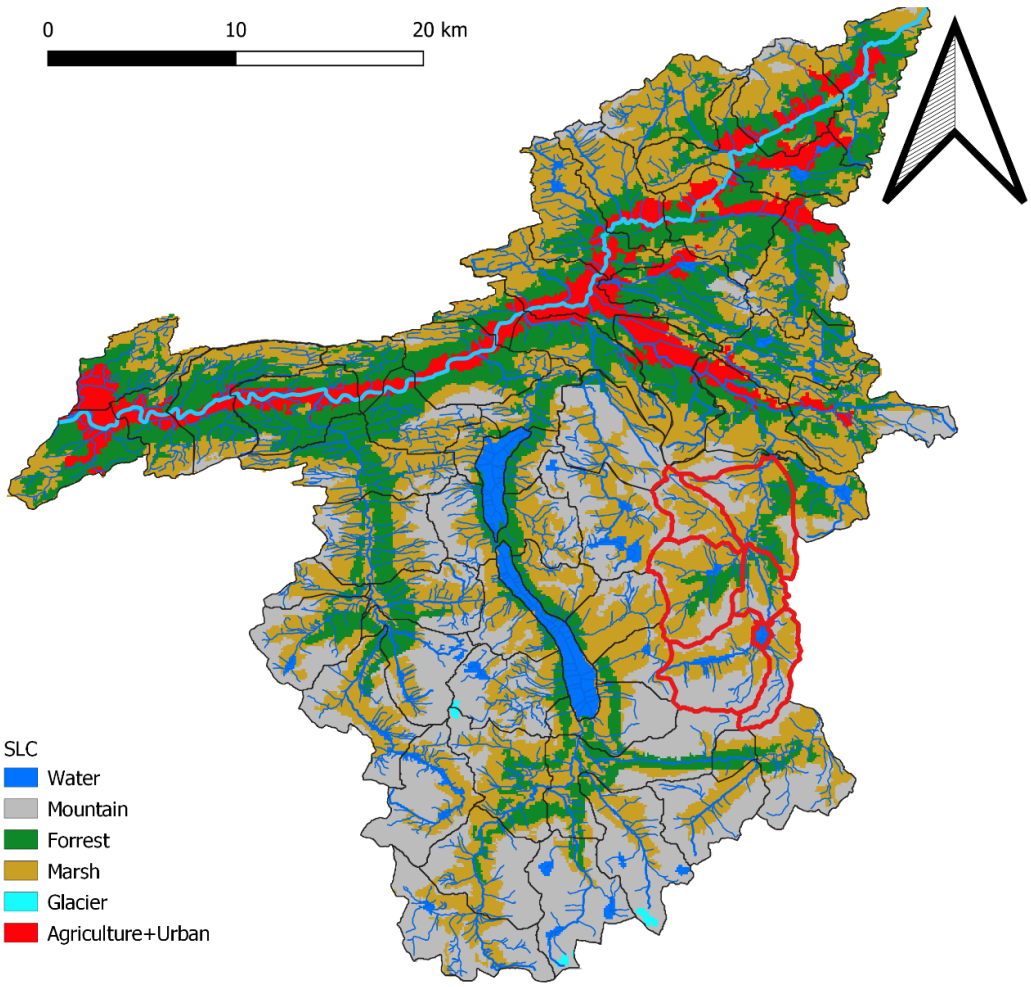


Figure 54: SLCs of the Surna catchment

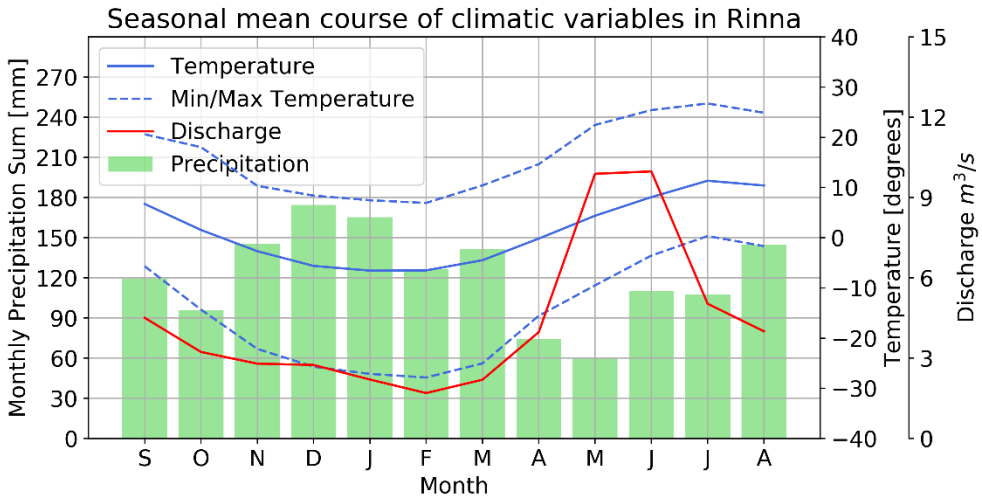


Figure 55: Seasonal mean course of the temperature, precipitation and discharge in a time span of 20 years (2000-2020) in Rinna. Data: seNorge2018 daily

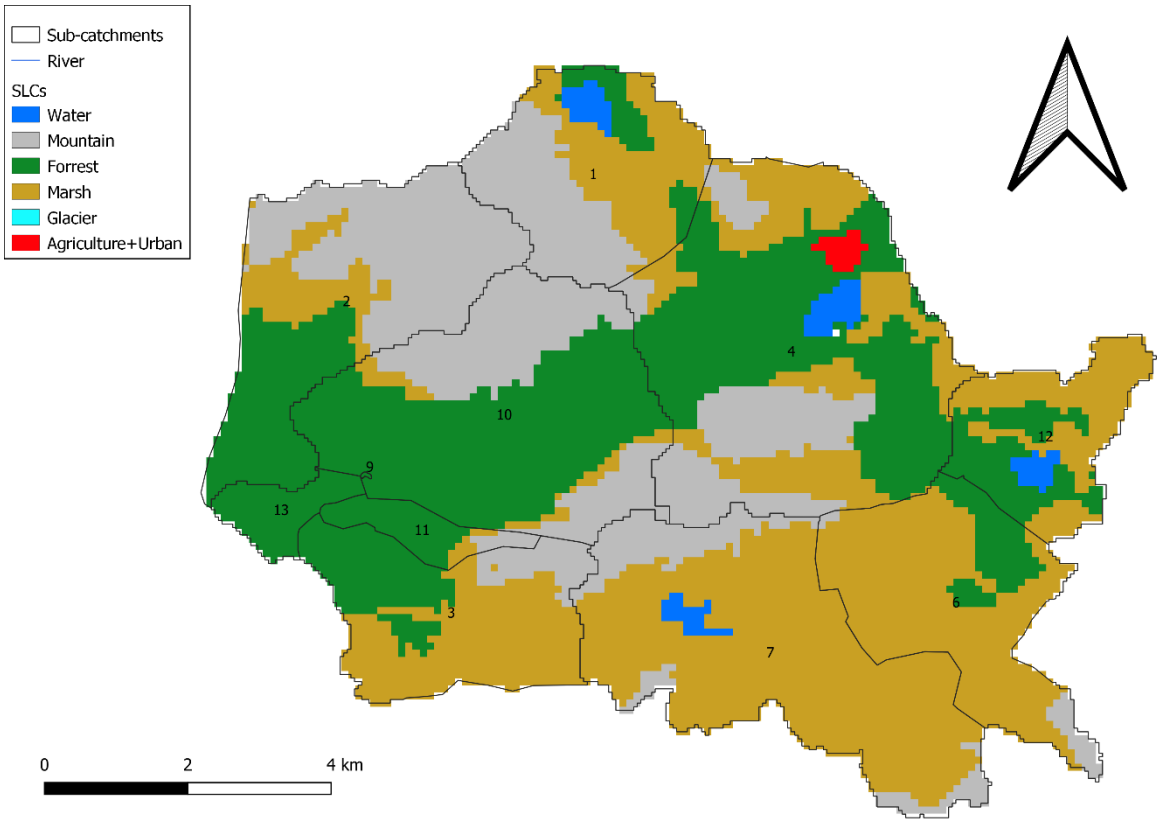


Figure 56: SLCs of the Grunnåi catchment

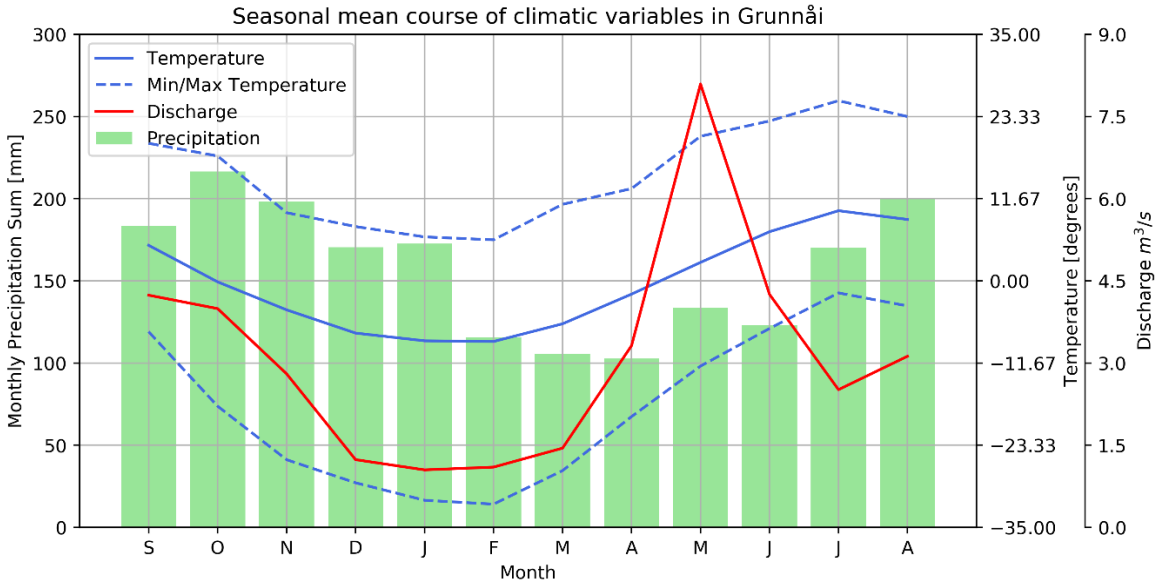


Figure 57: Seasonal mean course of the temperature, precipitation and discharge in a time span of 20 years (2000-2020) in Grunnåi. Data: seNorge2018 daily

B.2 Data origin

Table 18: Summary of applied datasets in this study and their access

	Source
seNorge2 hourly temperature	Lussana et al. (2016b) https://thredds.met.no/thredds/catalog/senorge/seNorge2 (10.12.2021)
NWP hourly temperature	Nipen et al. (2020) https://thredds.met.no/thredds/catalog/metpparchivev2 (10.12.2021)
seNorge2018 daily temperature	Lussana et al. (2019b) https://thredds.met.no/thredds/catalog/senorge/seNorge_2018/version_18.12/Archive/catalog.html (10.12.2021)
seNorge2 hourly precipitation	Lussana et al. (2016b) https://thredds.met.no/thredds/catalog/senorge/seNorge2 (10.12.2021)
NWP hourly precipitation	https://thredds.met.no/thredds/catalog/metpparchivev2 (10.12.2021)
seNorge2018 daily precipitation	Lussana et al. (2019b) https://thredds.met.no/thredds/catalog/senorge/seNorge_2018/version_18.12/Archive/catalog.html (10.12.2021)
Raw Radar precipitation	https://thredds.met.no/thredds/catalog/remotesensingradaraccr/catalog.html (10.12.2021)
KliNoGrid precipitation	Lussana et al. (2016a)
Discharge NVE	https://sildre.nve.no/map?params=1001&maxAge=-1&showDisused=true&x=422726&y=7310225&zoom=13&lang=en (10.12.2021)
Evaporation product MODIS	https://earthdata.nasa.gov/ (17.03.2022)
Land use	https://land.copernicus.eu/pan-european/corine-land-cover/clc-2012?tab=metadata (10.12.2021)
Geodata of lakes, rivers, location of gauges and hydropower in Norway	NVE data base https://nedlasting.nve.no/gis/
Catchment delineation	Schönfelder and Baclet (2022)
Digital elevation map	https://kartkatalog.geonorge.no/metadata/dtm-50/e25d0104-0858-4d06-bba8-d154514c11d2 (10.12.2021)

C Dataset comparison

C.1 Annual precipitation sum

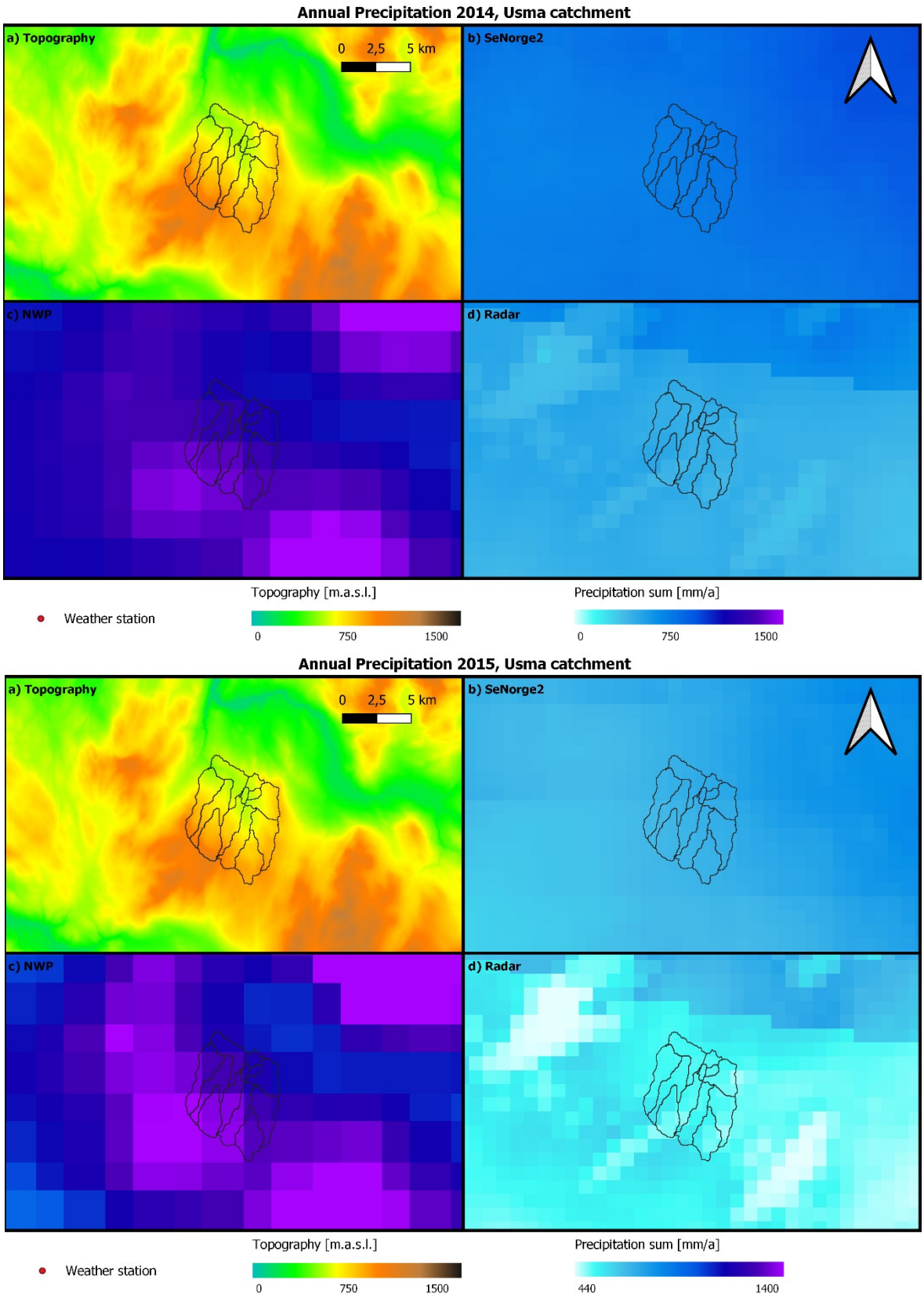


Figure 58: Annual precipitation sum of 2014 and 2015 over the Usma catchment

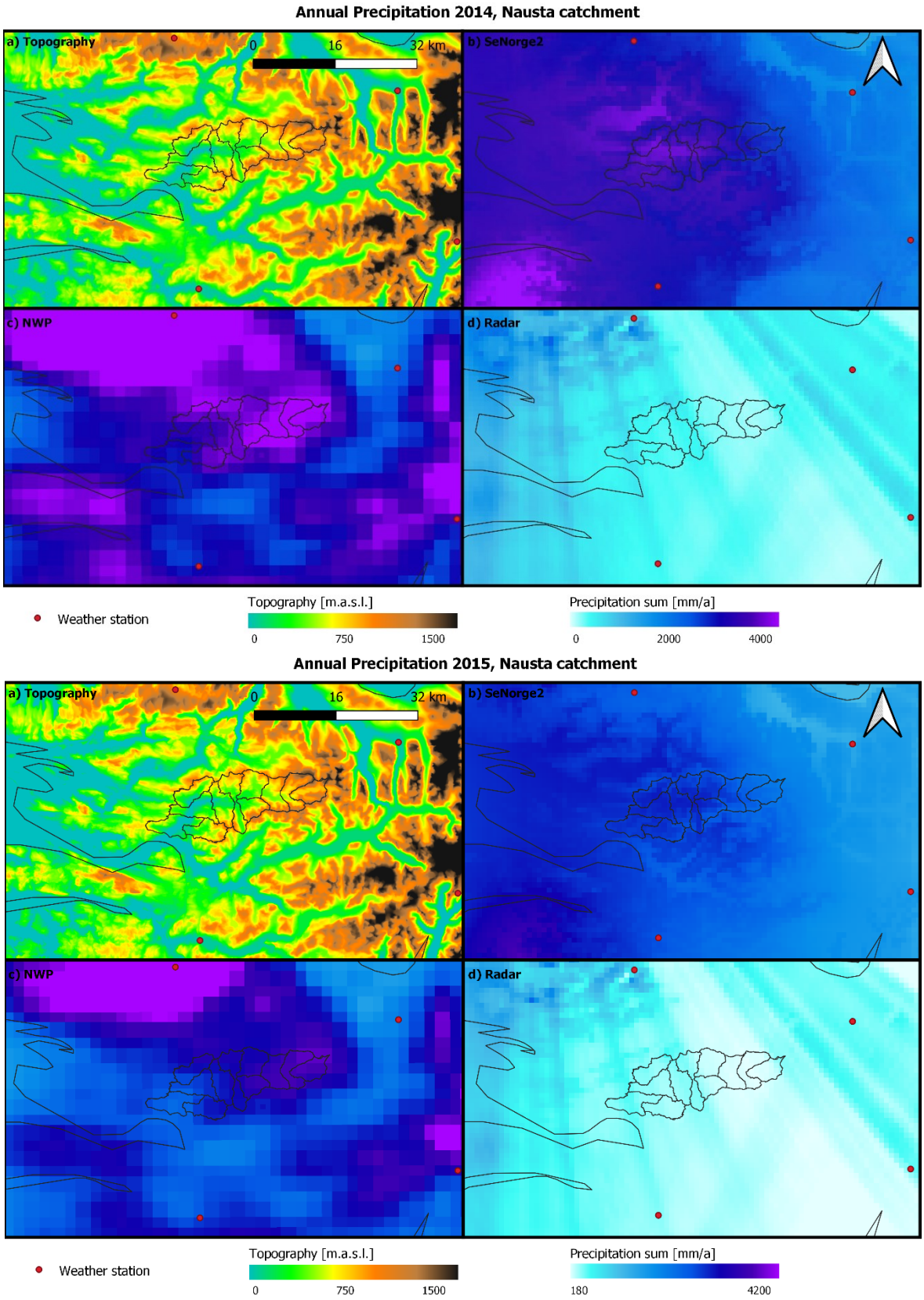


Figure 59: Annual precipitation sum of 2014 and 2015 over the Nausta catchment

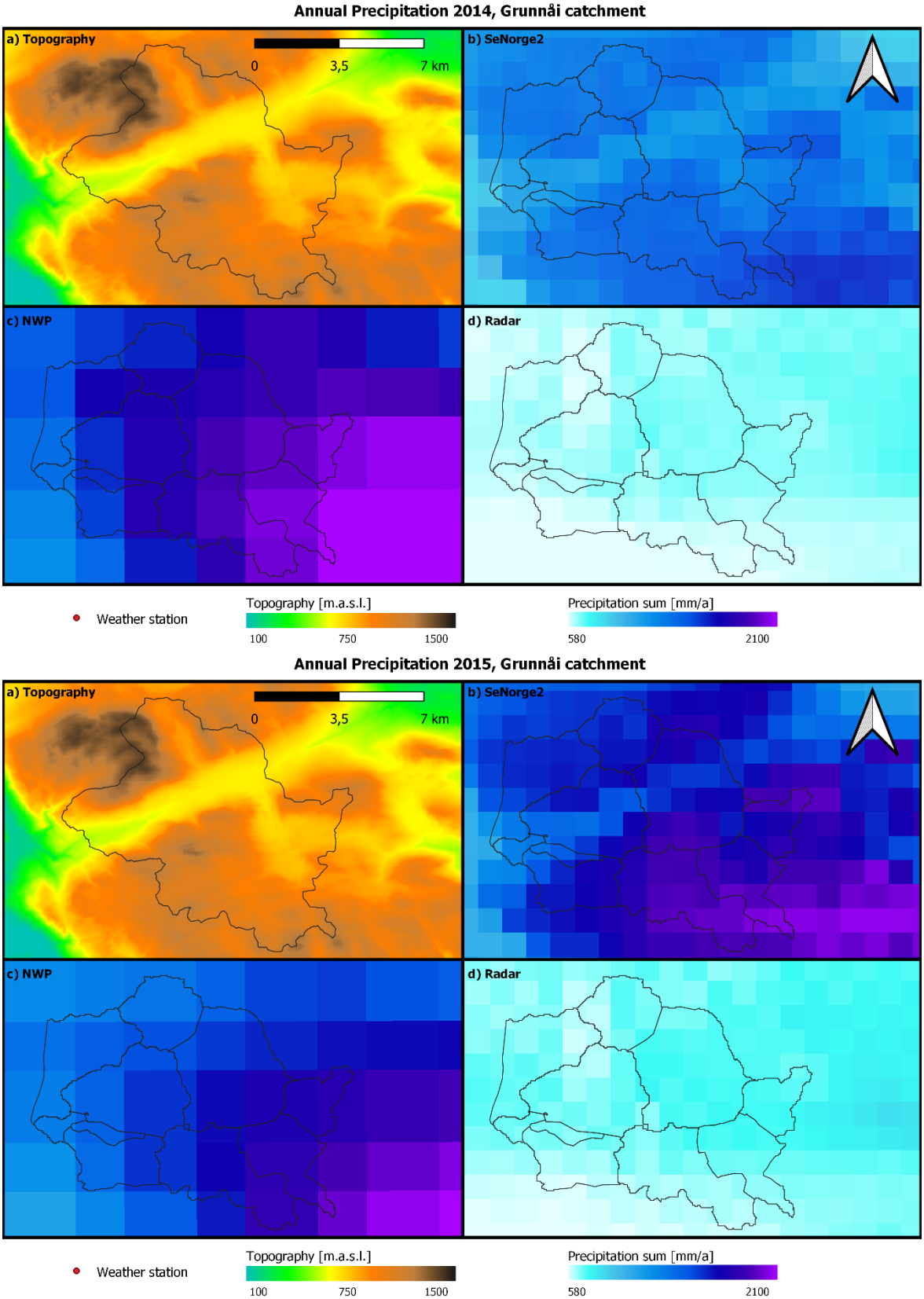


Figure 60: Annual precipitation sum of 2014 and 2015 over the Grunnåi catchment

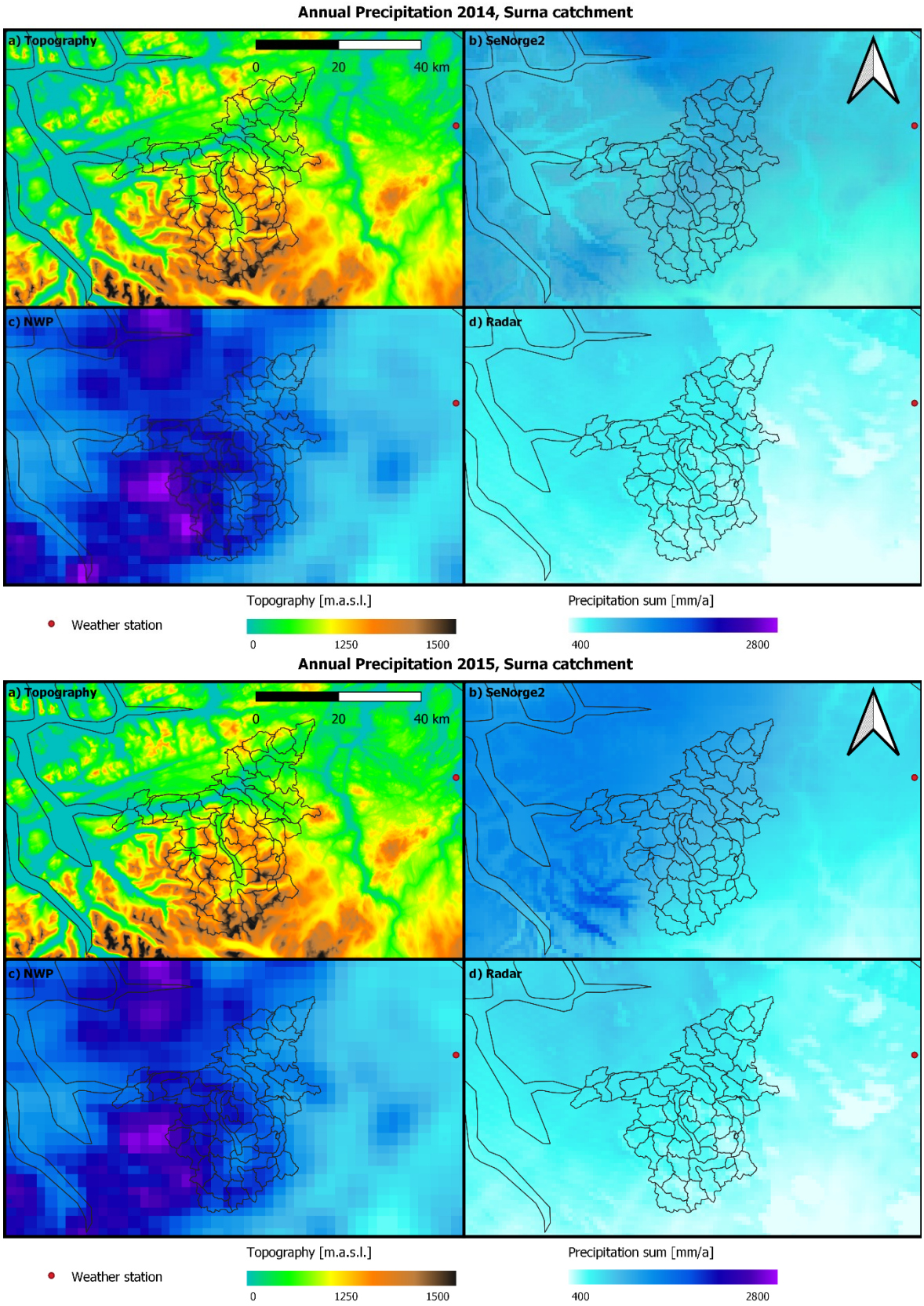


Figure 61: Annual precipitation sum of 2014 and 2015 over the Surna catchment

C.2 Correlation

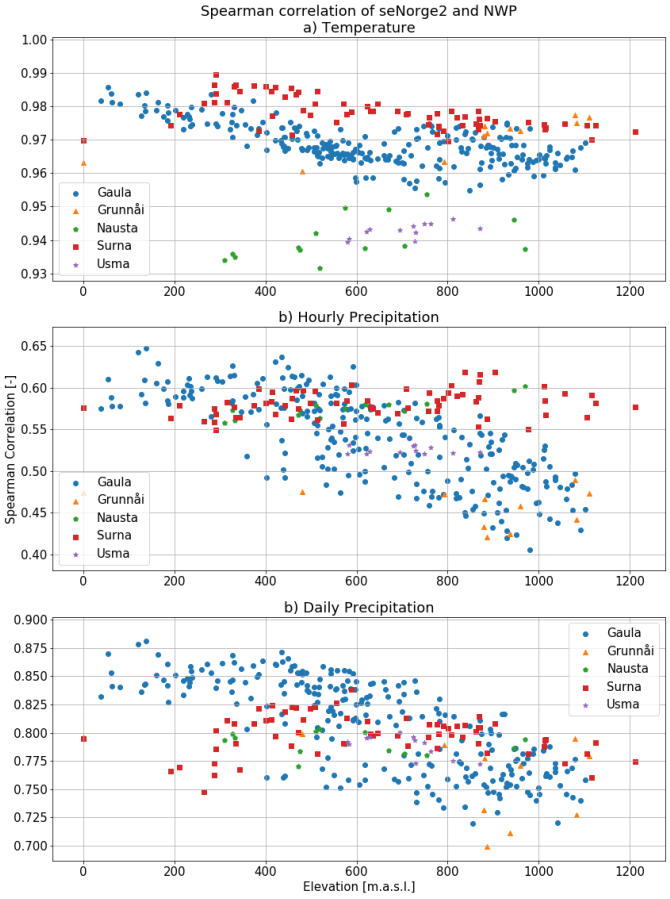


Figure 62: Spearman correlation seNorge2 and NWP of each sub-catchment, scattered over the mean elevation.

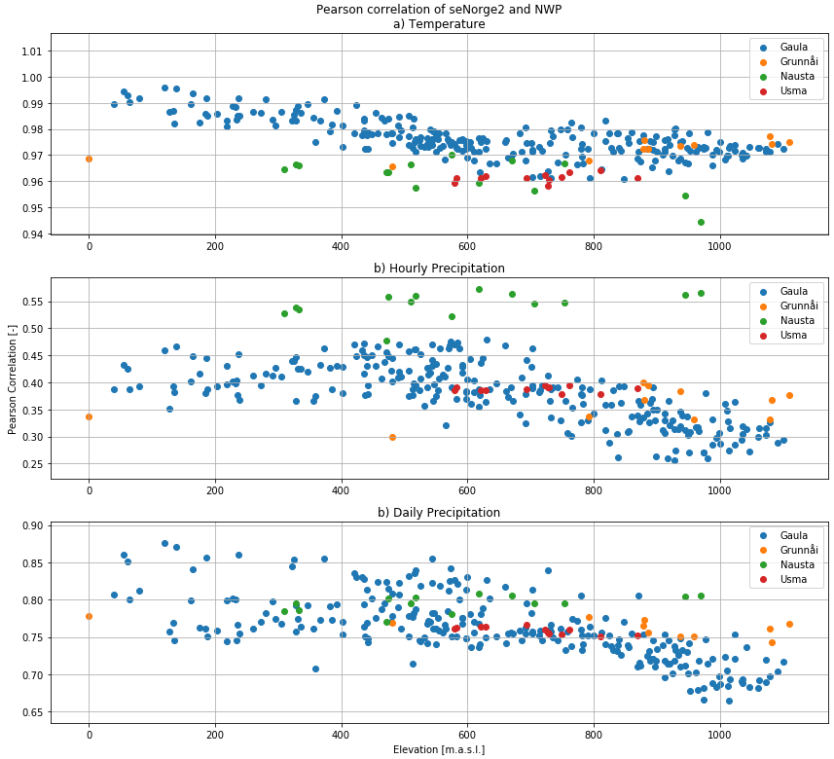


Figure 63: Pearson correlation seNorge2 and NWP of each sub-catchment, scattered over the mean elevation.

D Kirchner model

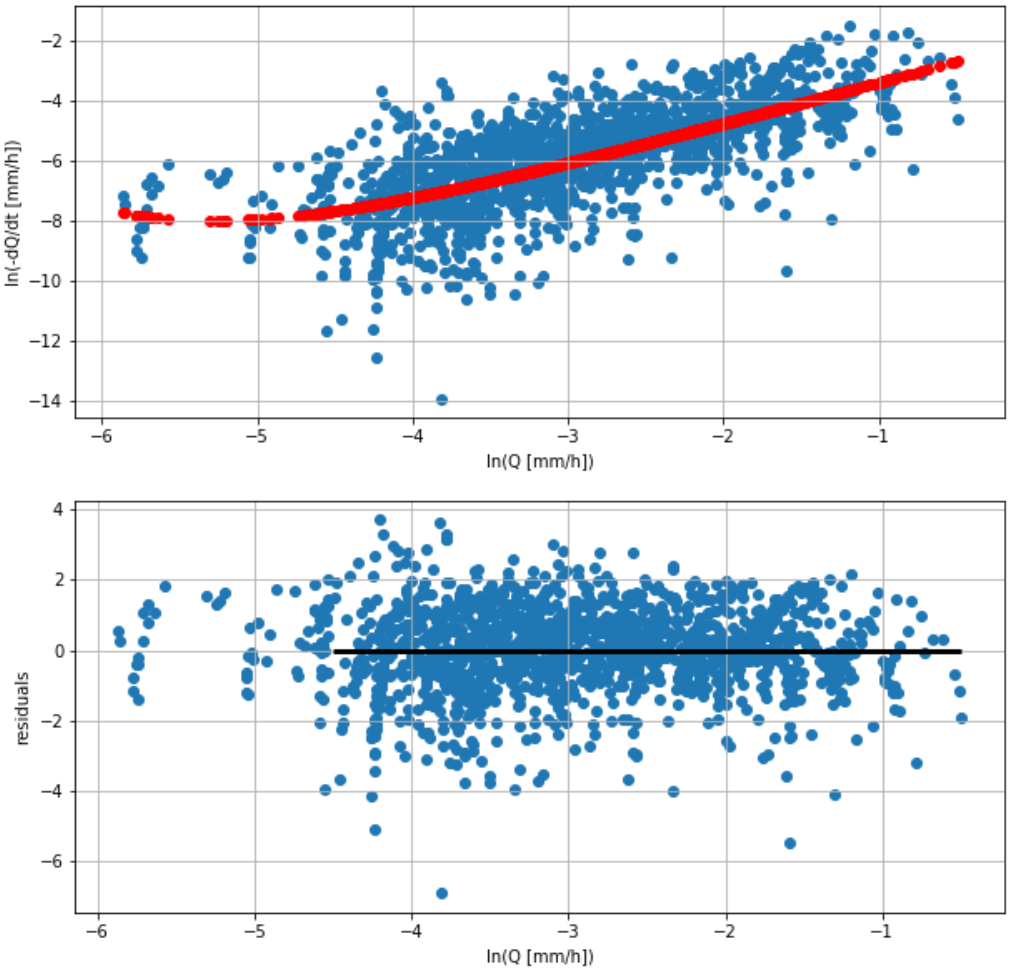


Figure 64: Sensitivity curve of Usma. Parameters: C_1: -1.97, C_2: 1.40, C_3: 0.007

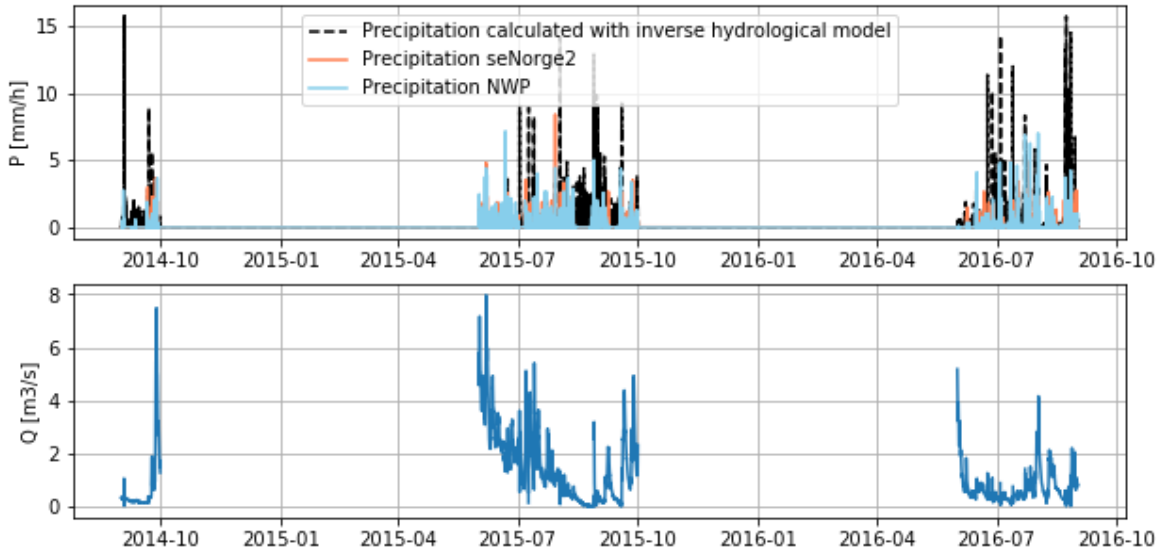


Figure 65: Simulated precipitation according to an inverse hydrological model

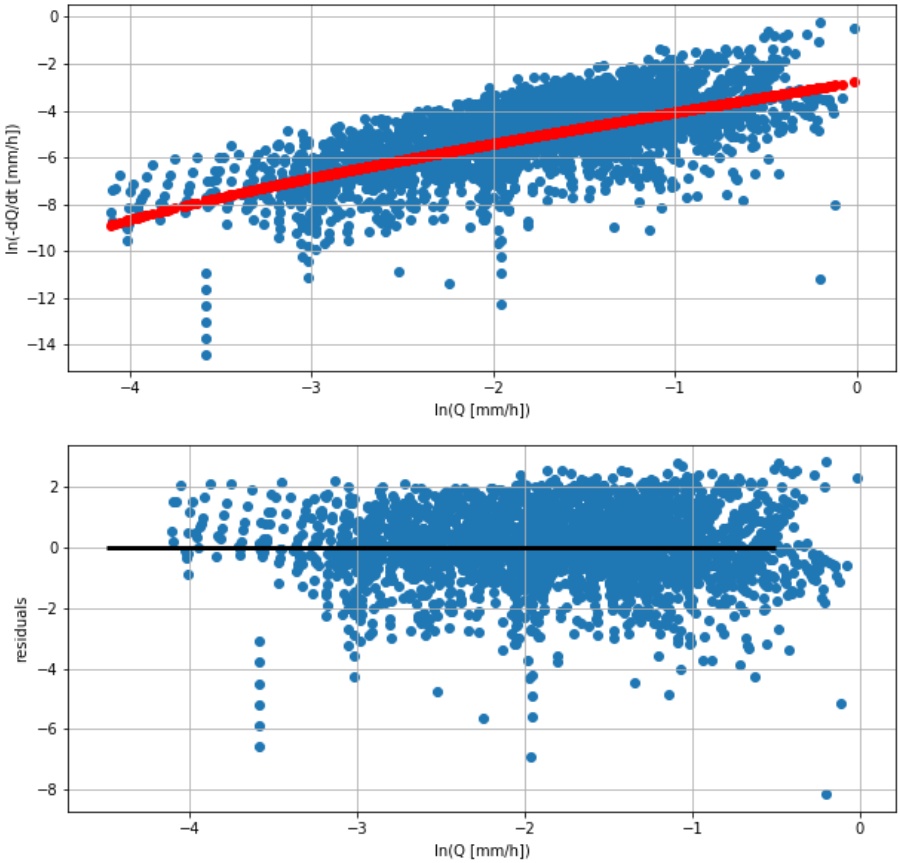


Figure 66: Sensitivity curve of Nausta. Parameters: C_1: -2.75, C_2: 1.27, C_3: -0.015

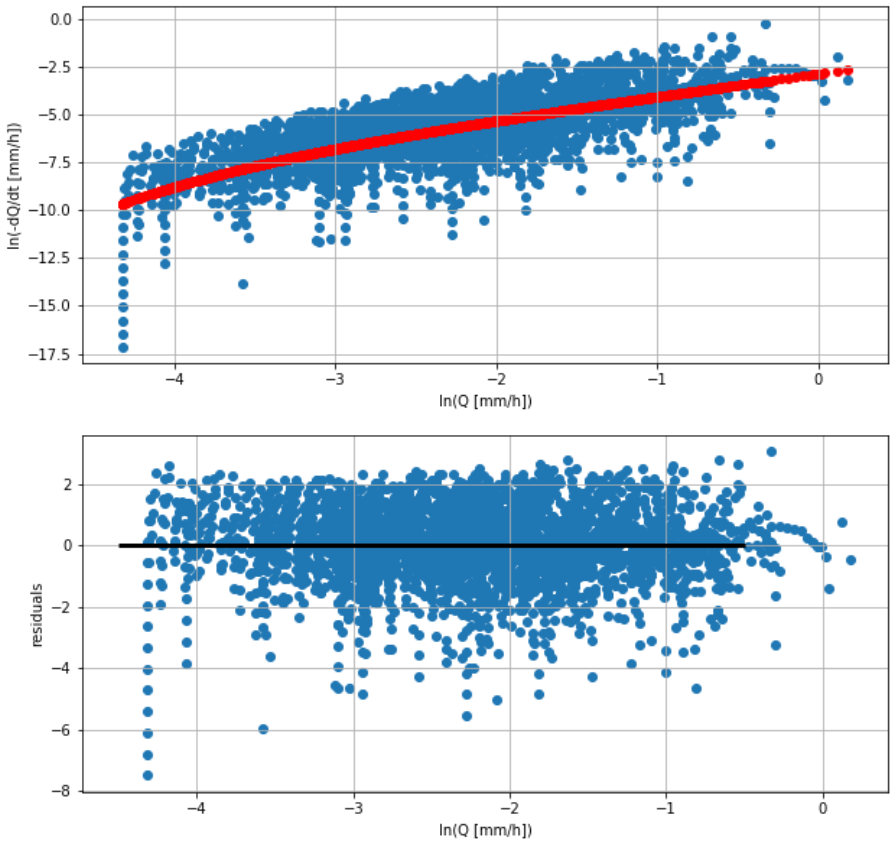


Figure 67: Sensitivity curve of Rinna. Parameters: C_1: -2.85, C_2: 1.15, C_3: -0.024

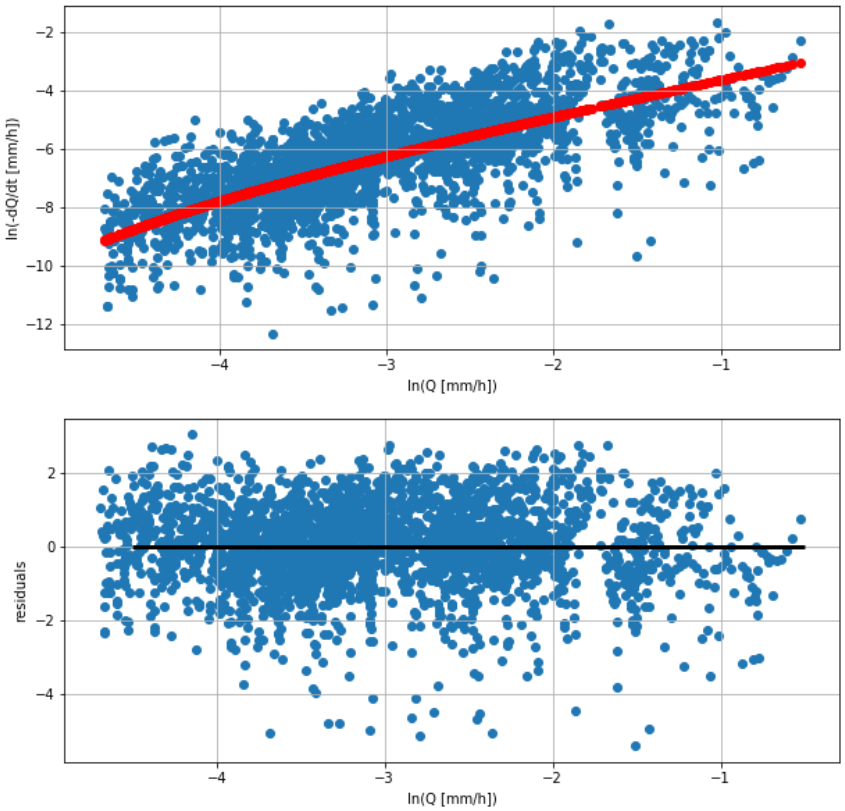


Figure 68: Sensitivity curve of Gaulfoss. Parameters: C_1: -2.36, C_2: 1.22, C_3: -0.009

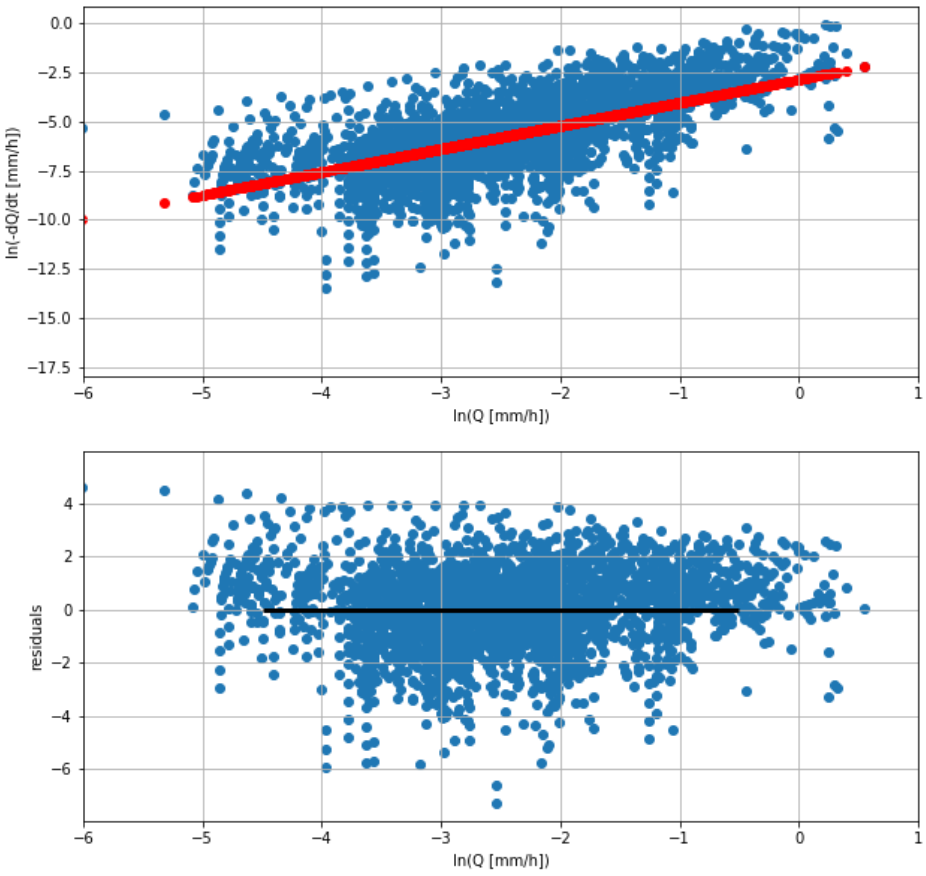


Figure 69: Sensitivity curve of Grunnåi. Parameters: C_1: -2.85, C_2: 178, C_3: 0.000003

E Calibration

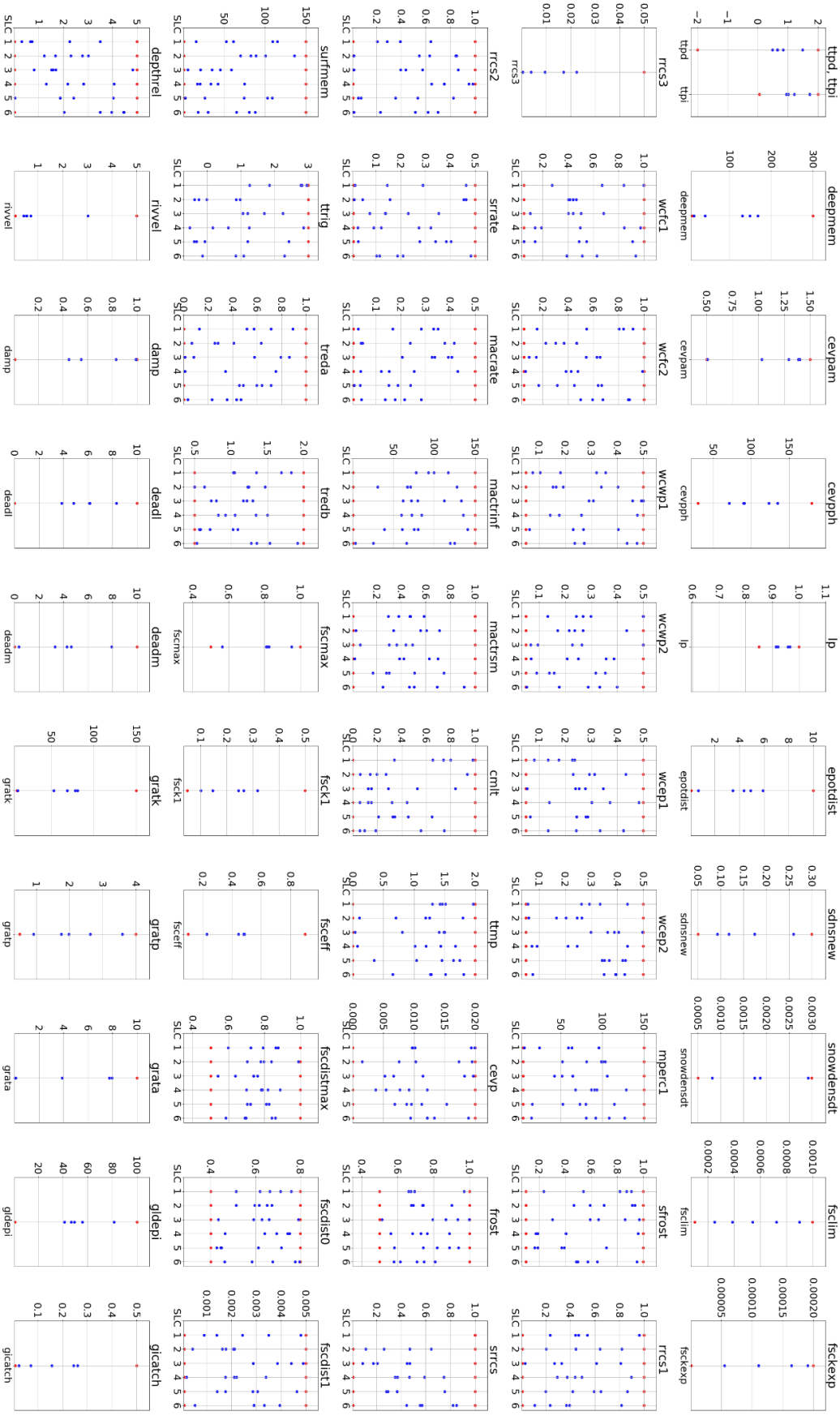


Figure 70: Parameter variability of the 20 best results using the DEMC method (number of simulations: 9000)

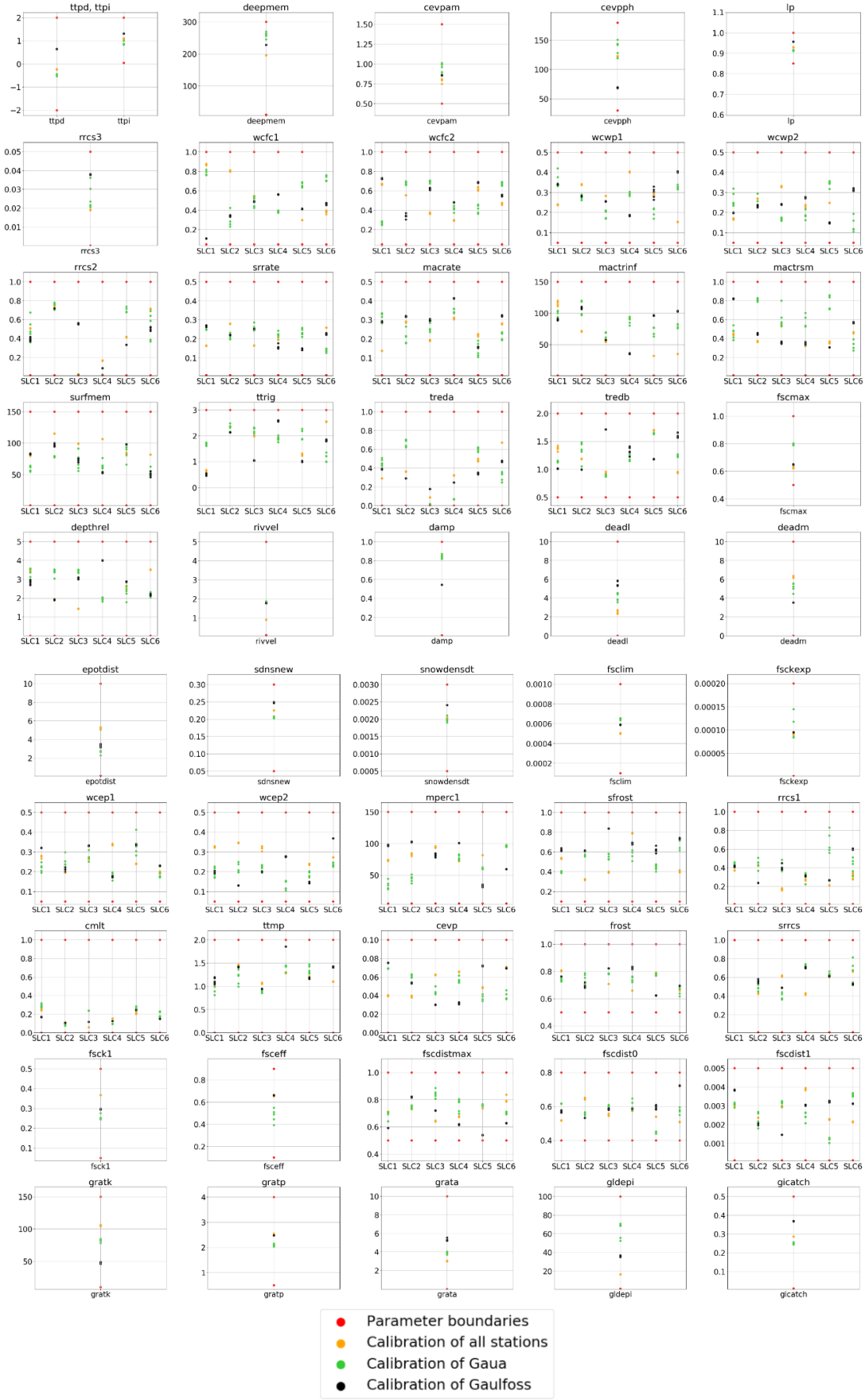


Figure 71: Comparison of parameter sets of different calibration methods: average KGE of all stations (yellow), calibration of Gaua (green) and Gaulfoss (black)

E.1 Stepwise calibration

Table 19: Results of stepwise calibration of the Nausta catchment

		seNorge2	corr. seNorge2	NWP	corr. NWP
Initial	**cal**	0.918	0.924	0.881	0.88
	val	0.725	0.758	0.833	0.844
Soil	**cal**	0.914	0.926	0.869	0.877
	val	0.726	0.756	0.826	0.843
Snow	**cal**	0.921	0.93	0.883	0.881
	val	0.725	0.743	0.825	0.845
Divers	**cal**	0.923	0.931	0.884	0.882
	val	0.727	0.747	0.825	0.846
Ilake	**cal**	0.929	0.935	0.868	0.869
	val	0.727	0.745	0.825	0.846
Olake	**cal**	0.923	0.931	0.884	0.882
	val	0.726	0.745	0.826	0.847
Soil	**cal**	0.9	0.928	0.884	0.882
	val	0.718	0.756	0.825	0.845
Snow	**cal**	0.921	0.93	0.884	0.883
	val	0.722	0.749	0.825	0.846
Divers	**cal**	0.921	0.929	0.884	0.883
	val	0.723	0.745	0.825	0.846
Ilake	**cal**	0.928	0.934	0.87	0.869
	val	0.722	0.745	0.825	0.846
Olake	**cal**	0.921	0.93	0.884	0.883
	val	0.721	0.744	0.825	0.846

Table 20: Results of stepwise calibration of the Usma catchment

		seNorge2	corr. seNorge2	NWP	corr. NWP	WS	corr. WS
Initial	**cal**	0.734	0.82	0.859	0.849	0.85	0.87
Soil	**cal**	0.73	0.847	0.864	0.853	0.844	0.868
Snow	**cal**	0.741	0.875	0.867	0.852	0.85	0.872
Rest	**cal**	0.759	0.88	0.867	0.854	0.855	0.876
Ilake	**cal**	0.759	0.88	0.868	0.854	0.855	0.876
Olake	**cal**	0.775	0.882	0.87	0.856	0.863	0.879
Soil	**cal**	0.789	0.88	0.871	0.854	0.857	0.88
Snow	**cal**	0.794	0.885	0.87	0.856	0.863	0.88
Divers	**cal**	0.794	0.888	0.873	0.859	0.863	0.881
Ilake	**cal**	0.794	0.888	0.873	0.859	0.863	0.881
Olake	**cal**	0.794	0.89	0.876	0.86	0.864	0.882

Table 21: Results of stepwise calibration of the Grunnåi catchment

		seNorge2	corr. seNorge2	NWP	corr. NWP
Initial	**cal**	0.822	0.845	0.794	0.809
	val	0.507	0.507	0.709	0.759
Soil	**cal**	0.821	0.820	0.795	0.812
	val	0.506	0.507	0.710	0.757
Snow	**cal**	0.838	0.838	0.798	0.811
	val	0.556	0.558	0.715	0.758
Divers	**cal**	0.845	0.818	0.799	0.817
	val	0.536	0.500	0.714	0.770
Ilake	**cal**	0.847	0.840	0.800	0.817
	val	0.533	0.555	0.717	0.772
Olake	**cal**	0.848	0.845	0.802	0.819
	val	0.526	0.541	0.715	0.770
Soil	**cal**	0.848	0.845	0.803	0.819
	val	0.527	0.537	0.715	0.769
Snow	**cal**	0.854	0.848	0.801	0.818
	val	0.557	0.559	0.715	0.770
Divers	**cal**	0.850	0.855	0.802	0.819
	val	0.546	0.547	0.718	0.773
Ilake	**cal**	0.851	0.855	0.802	0.819
	val	0.546	0.546	0.718	0.773
Olake	**cal**	0.854	0.856	0.803	0.819
	val	0.535	0.542	0.716	0.771

Table 22: Results of stepwise calibration of the Gaula catchment

		seNorge2	corr. seNorge2	NWP	corr. NWP
Initial	**cal**	0.685	0.713	0.708	0.722
	val	0.707	0.794	0.570	0.586
Soil	**cal**	0.652	0.745	0.720	0.733
	val	0.693	0.794	0.581	0.587
Snow	**cal**	0.694	0.757	0.722	0.718
	val	0.707	0.789	0.571	0.571
Divers	**cal**	0.696	0.760	0.724	0.737
	val	0.720	0.791	0.571	0.585
Ilake	**cal**	0.696	0.761	0.724	0.737
	val	0.720	0.791	0.57	0.570
Olake	**cal**	0.697	0.762	0.726	0.740
	val	0.720	0.793	0.570	0.587

Table 23: Results stepwise calibration Surna catchment

		seNorge2	corr. seNorge2	NWP	corr. NWP
Initial	**cal**	0.709	0.866	0.818	0.824
	val	0.505	0.703	0.652	0.675
Soil	**cal**	0.715	0.834	0.818	0.826
	val	0.505	0.697	0.651	0.669
Snow	**cal**	0.715	0.865	0.824	0.821
	val	0.509	0.692	0.645	0.683
Divers	**cal**	0.720	0.874	0.826	0.826
	val	0.505	0.700	0.644	0.673
Ilake	**cal**	0.719	0.875	0.828	0.827
	val	0.505	0.697	0.643	0.672
Olake	**cal**	0.722	0.876	0.820	0.828
	val	0.499	0.695	0.652	0.672
Soil	**cal**	0.714	0.849	0.831	0.826
	val	0.498	0.695	0.633	0.678
Snow	**cal**	0.728	0.881	0.832	0.825
	val	0.507	0.715	0.633	0.690
Divers	**cal**	0.731	0.881	0.832	0.829
	val	0.511	0.713	0.633	0.679
Ilake	**cal**	0.731	0.882	0.832	0.830
	val	0.511	0.712	0.633	0.680
Olake	**cal**	0.731	0.882	0.831	0.830
	val	0.506	0.711	0.633	0.680

E.2 Low Flow calibration

This chapter explains the method and summarizes the results of trying to improve low flows in the model. This was done in the Nausta catchments and results applied into another catchment (Usma). In the first step, the catchment was only calibrated for the lowest 20 % of discharge. This resulted in a poor performance as well as a not suitable CDF function.

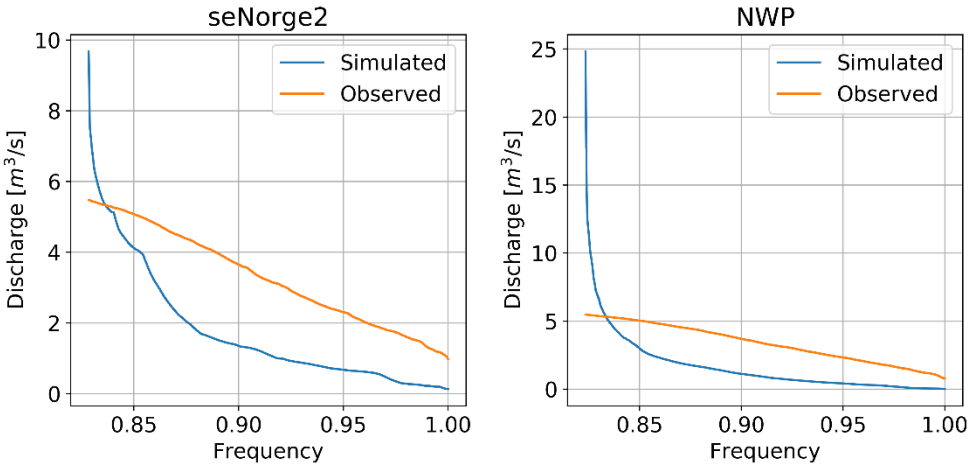


Figure 72: CDF of calibrated low flows in Nausta, with seNorge2 (left) and NWP (right)

The next strategy was changing the objective function to test whether using the mean squared error (HYPE: MNR) or relative bias (HYPE: MRE) would improve the low flows of the model. This was applied for the Q80 calibration, and the results compared.

Table 24: Results of testing different objective functions

dataset	Obj. function	KGE	RMSE	relative bias
seNorge2	KGE	0.305	0.330	-0.180
	MRE	-2.947	1.172	0.0007
	MNR	0.199	0.358	-0.160
NWP	KGE	0.209	0.371	-0.215
	MRE	-2.998	1.189	-0.00006
	MNR	0.205	0.350	-0.115

The CDF of the RMSE calibration looked similar to the one from KGE, the relative bias deteriorated the results. It can therefore be assumed that the poor performance of low flows is not caused by a not suitable objective function.

To get an overview over parameters that have an influence on low flows, a simple algorithm was implemented, that manually changed each parameter. For that, every parameter was set to his lowest, highest and average value and the model was executed. The KGE was then calculated and compared to the KGE of the original model. Parameter with high changes were considered sensitive to low flows. This sensitivity analysis was done for Q80 discharges.

The results of each parameter are summarized in Table 25.

Table 25: Results of the sensitivity analysis of low flow calibration

	Minimum value	Maximum value	Mean value	KGE change with max [%]	KGE change with min [%]	KGE change with mean [%]
ttpd	-1.98	1.98	0.00	-25.4	25.4	4.7
ttpi	0.05	1.98	1.02	12.6	-1.0	3.8
deepmem	10.10	297.00	153.55	-0.3	2.1	0.0
cevpam	0.51	1.49	1.00	0.0	0.0	0.0
cevpph	30.30	178.20	104.25	0.0	0.0	0.0
lp	0.86	0.99	0.92	0.0	0.0	0.0
epotdist	0.10	9.90	5.00	0.0	-0.1	0.0
sdnsnew	0.05	0.30	0.17	0.1	-0.1	0.0
snowdensdt	0.00	0.00	0.00	0.1	-0.7	-0.1
fsclim	0.00	0.00	0.00	0.0	0.0	0.0
fsckexp	0.00	0.00	0.00	0.1	-18.1	0.0
wcfc1	0.05	0.99	0.52	0.0	0.1	0.0
wcfc2	0.05	0.99	0.52	0.0	0.0	0.0
wcfc3	0.05	0.99	0.52	0.1	-0.1	0.0
wcwp1	0.05	0.50	0.27	0.0	0.0	0.0
wcwp2	0.05	0.50	0.27	0.0	0.0	0.0
wcwp3	0.05	0.50	0.27	0.1	0.0	0.0
wcep1	0.05	0.50	0.27	0.0	0.0	0.0
wcep2	0.05	0.50	0.27	0.0	0.0	0.0
wcep3	0.05	0.50	0.27	0.0	0.0	0.0
mperc1	5.05	148.50	76.78	0.0	0.0	0.0
mperc2	5.05	148.50	76.78	0.0	0.0	0.0
sfrost	0.10	0.99	0.55	0.0	0.0	0.0
rrcs1	0.01	0.99	0.50	0.0	0.0	0.0
rrcs2	0.01	0.99	0.50	-8.1	40.8	-8.0
rrcs3	0.01	0.99	0.50	0.0	0.0	0.0
srrate	0.01	0.50	0.25	0.0	0.0	0.0
macrate	0.01	0.50	0.25	0.0	0.0	0.0
mactrinf	1.01	148.50	74.76	0.0	-3.5	0.0
mactrsm	0.01	0.99	0.50	0.0	0.0	0.0
cmlt	0.01	0.49	0.25	-176.9	-84.5	-57.4
ttmp	0.01	1.98	0.99	-35.5	-36.1	-1.2
cevp	0.00	0.10	0.05	0.0	0.0	0.0
frost	0.51	0.99	0.75	0.0	0.0	0.0
srrcs	0.01	0.99	0.50	0.0	0.0	0.0
surfmem	1.01	148.50	74.76	0.0	0.2	0.0
ttrig	-0.99	2.97	0.99	-1.5	0.2	0.2
treda	0.00	0.99	0.50	-0.1	0.0	-0.2
tredb	0.51	1.98	1.24	-0.2	-0.3	-0.3
fscmax	0.51	0.99	0.75	-1.4	-69.4	-14.4
fsck1	0.05	0.50	0.27	0.8	-34.7	-10.3
fsceff	0.10	0.89	0.50	-25.8	-6.3	1.1
fscdistmax	0.51	0.99	0.75	0.0	0.3	0.0
fscdist0	0.40	0.79	0.60	-9.3	0.6	-4.8

fscdist1	0.00	0.00	0.00	0.0	0.0	0.0
depthrel	0.01	4.95	2.48	-0.2	0.0	-0.1
rivvel	0.10	4.95	2.53	16.4	-46.8	-66.4
damp	0.01	0.99	0.50	10.6	-23.7	-7.8
deadl	0.00	9.90	4.95	0.0	0.0	0.0
deadm	0.00	9.90	4.95	0.0	0.0	0.0
gratk	10.10	148.50	79.30	-20.1	-9.9	-17.9
gratp	0.51	3.96	2.23	1.5	-33.0	2.7
grata	0.00	9.90	4.95	-39.5	9.4	-29.5
gldepi	1.01	99.00	50.01	0.0	0.0	0.0
gicatch	0.01	0.50	0.25	0.0	0.0	0.0

This analysis showed than not all of the parameters have an influence on low flows. According to the changes in the KGE the parameters: *ttpd*, *ttpi, deepmem, snowdensdt, fsckexp, rrcs2, mactrinf, cmelt, ttmp, ttrig, treda, tredb, fscmax, fsck1, fsceff, fscdistmax, fscdist0, depthrel, rivvel, damp, graktk, gratp, grata* have the highest influence.

The sensitivity analysis was repeated using RMSE instead of KGE. The sensitive parameters remained similar. The sensitive parameters were afterwards used as the only optimization parameters. Using only them did only increase the performance slightly.

Because a good performance was not achieved even with a lower number of parameters, lead to the assumption that maybe the defined boundaries of the parameters were wrong. That's why in the next step, a GLUE analysis was applied. The boundaries were increased by decreasing the minimum and increasing the maximum. The parameters were calibrated using Monte-Carlo with 100,000 simulations. Afterwards, the best 100 parameter sets were compared. It became obvious that the parameters vary too much and did not converge at all. Except for the river velocity.

Figure 73: Results of GLUE analysis of low flow calibration in Nausta

Because the model was not able to converge against a certain value, it was decided to manually change the boundaries, according to assumptions of the parameter behaviour. The `optpar` file with the changed boundary values can be found in the digital appendix. Especially recession parameters were decreased to improve the performance. It is possible that these values are also timestep dependant and should have been decreased in the first place. The initial Q80 calibration of seNorge2 had a KGE of 0.339, with increased boundaries, a KGE of 0.542 and with the manually changed boundaries 0.539. NWP was originally 0.023, with a higher parameter space 0.162 and finally 0.091. It seems like NWP is not able to create as good of a fit as NWP. This might be caused by the calibration method, because only specific periods are portrayed. If NWP has another timing than the other product, a peak might be measured at a wrong time step.

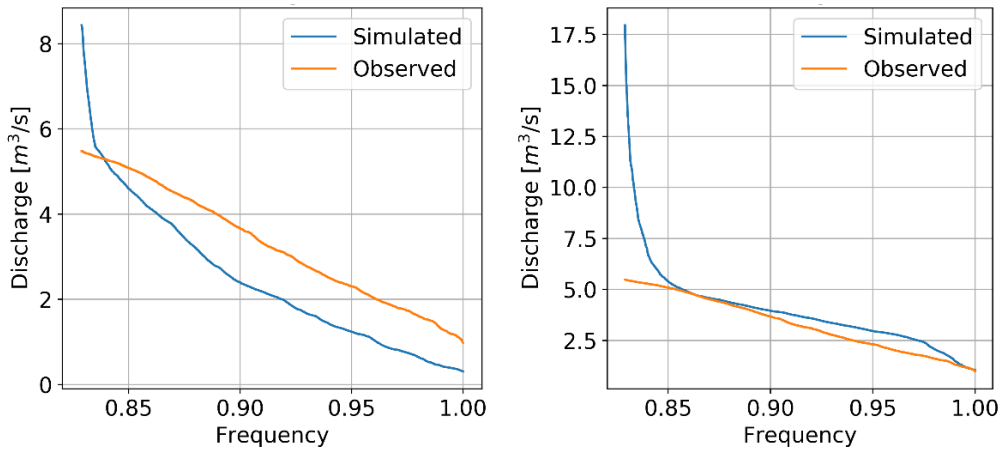


Figure 74: CDF of the final calibrated Q80 discharges in Nausta

A Q80 calibration was also done in Usma, also resulting in a worse performance for the NWP product than seNorge2. In Saldaña Espinoza (2022), the low flows were described better, without even focusing on them. This might be caused by the extra calibration of lake levels, that automatically calibrate storage and recession curves of a catchment. When using a daily timestep, timing issues will decrease. A possible improvement might be to calibrate recession and lake parameters by only calibrating the recession curves of a hydrograph and then calibrate other parameters in the next step.

This analysis resulted in no promising solutions. Although it must be assumed that the parameter values have an influence, there was no optimal solution found. However, seNorge2 seemed to create better results than NWP. Furthermore, sensitive parameters were found and the need of an extra calibration step discussed.

F Model results

Table 26: Summarized model results of Usma

		SeNorge2	SeNorge2 corr	NWP	NWP corr	WS	WS corr
Correction	**Pcurain**	-	0.216	-	-0.340	-	0.135
	Pcusnow	-	0.466	-	-0.028	-	0.123
Difference calibrations	**KGE improvement**	0.06	0.07	0.02	0.01	0.01	0.01
	NSE improvement	0.11	0.14	0.03	0.03	0.03	0.02
Performance	**KGE**	0.780	0.885	0.874	0.859	0.857	0.879
	KGE summer	0.812	0.900	0.876	0.864	0.851	0.885
	KGE winter	0.418	0.657	0.723	0.604	0.469	0.588
	NSE	0.669	0.789	0.752	0.728	0.736	0.764
	NSE summer	0.650	0.800	0.752	0.743	0.738	0.780
	NSE winter	0.418	0.526	0.498	0.390	0.459	0.453
CDF	**Low flows***	(g)	(o)	(su)	(o)	(g)	(g)
	Low flows summer*	(so)	(so)	(u)	(o)	(o)	(o)
	Low flows winter*	(u)	(o)	(su)	(o)	(su)	(g)
	Peaks*	(o)	(o)	(o)	(o)	(o)	(o)
	Peaks summer*	(o)	(o)	(o)	(o)	(o)	(o)
	Peaks winter*	(u)	(u)	(u)	(u)	(u)	(u)
Correlation	**r^2 all**	0.69	0.8	0.77	0.74	0.75	0.78
	r^2 summer	0.67	0.81	0.77	0.77	0.76	0.8
	r^2 winter	0.48	0.81	0.77	0.77	0.76	0.8

* underestimated (u), slightly underestimated (su), in good agreement (g), slightly overestimated (so) and overestimated (o)

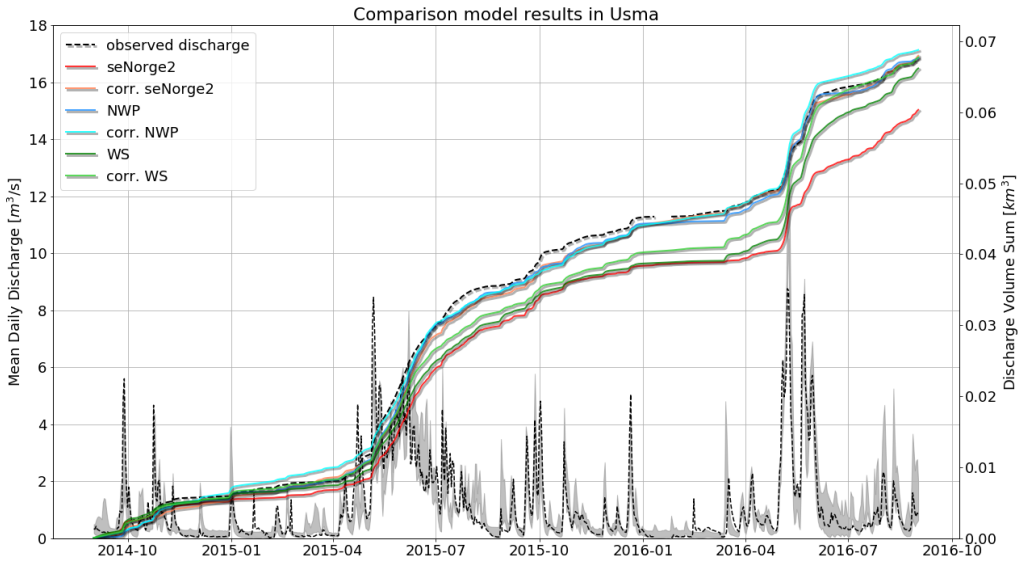


Figure 75: Comparison modelling results of Usma with different datasets. The time series was averaged to a daily timestep. Grey area: simulated discharge from the different products

Table 27: Summarized model results of Grunnåi

		SeNorge2	SeNorge2 corr	NWP	NWP corr	WS**	WS corr**
Correction	**Pcurain**	-	-0.031	-	0.217	-	0.465
	Pcusnow	-	0.284	-	-0.051	-	0.419
Difference calibrations	**KGE improvement**	0.03	0.01	0.01	0.01	-	-
	NSE improvement	0.05	0.02	0.03	0.02	-	-
Performance	**KGE**	0.852	0.843	0.785	0.795	0.636	0.840
	KGE summer	0.826	0.810	0.729	0.780	0.729	0.774
	KGE winter	0.780	0.789	0.740	0.684	0.361	0.701
	NSE	0.703	0.700	0.573	0.592	0.612	0.685
	NSE summer	0.695	0.681	0.543	0.580	0.643	0.691
	NSE winter	0.640	0.656	0.560	0.508	0.367	0.554
CDF	**Low flows**	(su)	(su)	(su)	(su)	(u)	(u)
	Low flows summer	(su)	(su)	(su)	(u)	(u)	(su)
	Low flows winter	(u)	(u)	(u)	(u)	(u)	(u)
	Peaks	(u)	(u)	(u)	(u)	(u)	(su)
	Peaks summer	(u)	(u)	(u)	(u)	(u)	(su)
	Peaks winter	(g)	(g)	(su)	(su)	(u)	(g)
Correlation	**r^2 all**	0.74	0.73	0.62	0.65	0.65	0.73
	r^2 summer	0.72	0.73	0.68	0.7	0.51	0.64
	r^2 winter	0.72	0.71	0.58	0.62	0.65	0.73

* underestimated (u), slightly underestimated (su), in good agreement (g), slightly overestimated (so) and overestimated (o)

** Daily timestep

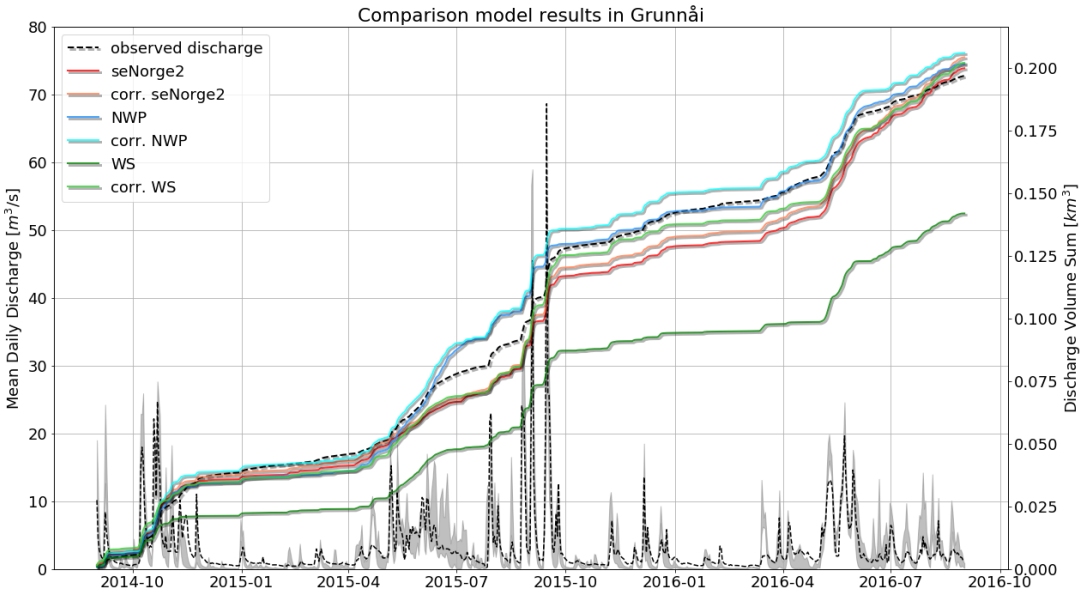


Figure 76: Comparison modelling results of Grunnåi with different datasets. The time series was averaged to a daily timestep. Grey area: simulated discharge from the different products

Table 28: Summarized model results of Surna

		SeNorge2	SeNorge2 corr	NWP	NWP corr
Correction	**pcurain**	-	0.481	-	-0.224
	pcusnow	-	0.249	-	0.279
Difference calibrations	**KGE improvement**	0,02	0,02	0,01	0,01
	NSE improvement	0,07	0,03	0,03	0,02
Performance	**KGE**	0,757	0,895	0,713	0,702
	KGE summer	0,801	0,899	0,698	0,682
	KGE winter	0,497	0,676	0,500	0,528
	NSE	0,703	0,799	0,639	0,629
	NSE summer	0,677	0,798	0,627	0,598
	NSE winter	0,193	0,364	-0,094	-0,017
CDF	**Low flows***	(u)	(su)	(u)	(su)
	Low flows summer*	(u)	(su)	(u)	(u)
	Low flows winter*	(u)	(u)	(u)	(u)
	Peaks*	(u)	(u)	(su)	(u)
	Peaks summer*	(u)	(u)	(u)	(u)
	Peaks winter*	(u)	(u)	(su)	(u)
Correlation	**r^2 all**	0,75	0,82	0,77	0,77
	r^2 summer	0,45	0,57	0,4	0,41
	r^2 winter	0,73	0,81	0,78	0,77

* underestimated (u), slightly underestimated (su), in good agreement (g), slightly overestimated (so) and overestimated (o)

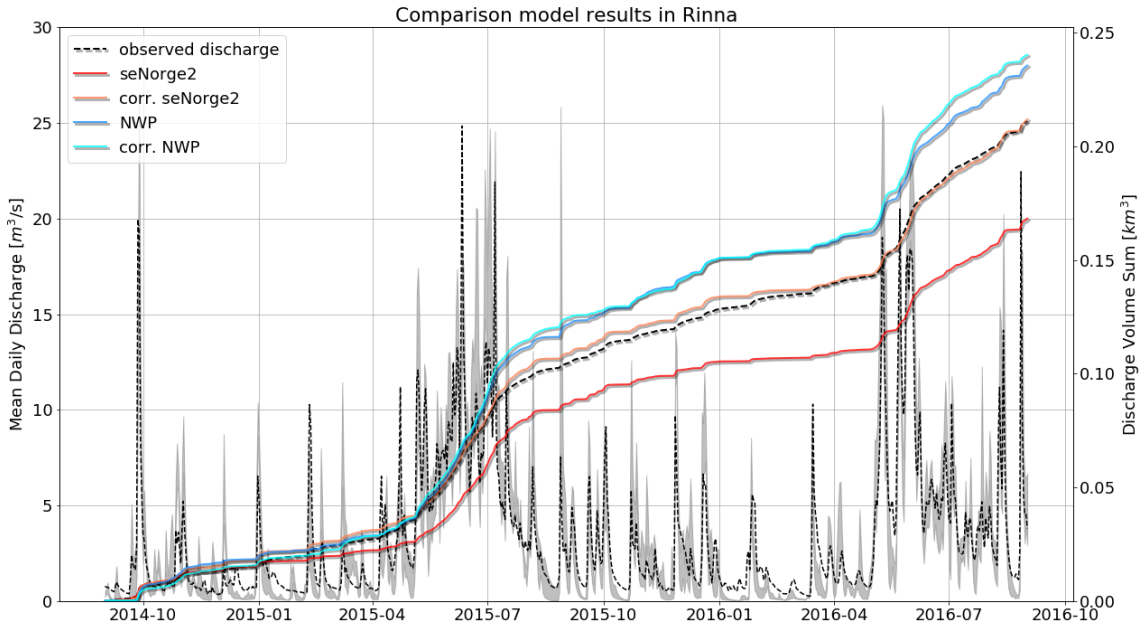


Figure 77: Comparison modelling results of Rinna with different datasets. The time series was averaged to a daily timestep. Grey area: simulated discharge from the different products

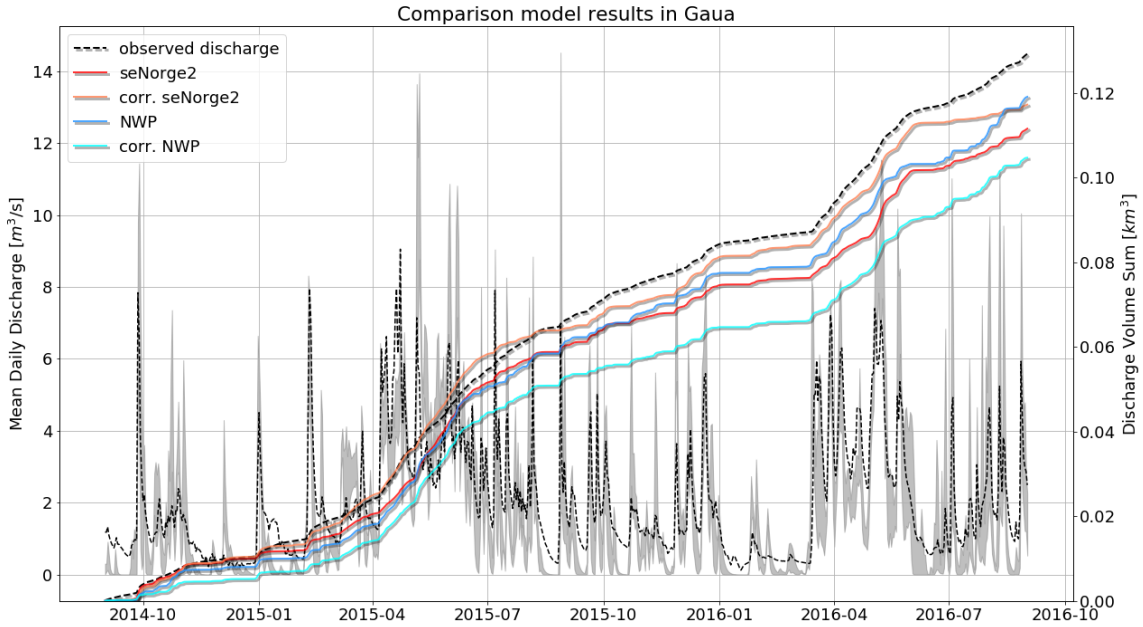


Figure 78: Comparison modelling results of Gaua (Gaula catchment) with different datasets. The time series was averaged to a daily timestep. Grey area: simulated discharge from the different products

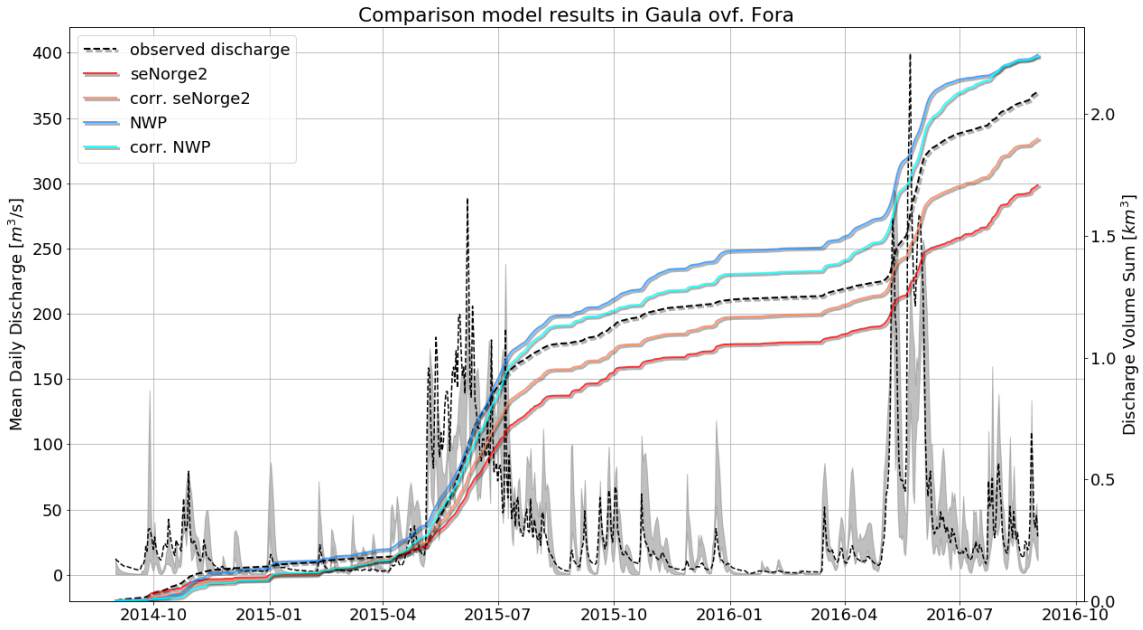


Figure 79: Comparison modelling results of Gaula ovf. Fora (Gaula catchment) with different datasets. The time series was averaged to a daily timestep. Grey area: simulated discharge from the different products

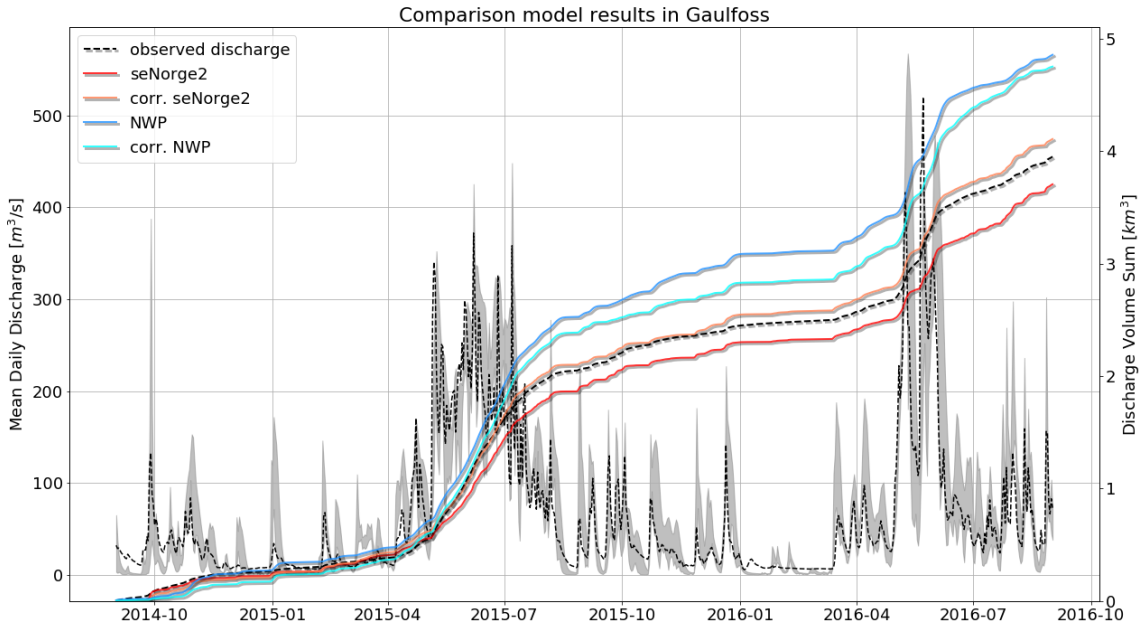


Figure 80: Comparison modelling results of Gaulfoss (Gaula catchment) with different datasets. The time series was averaged to a daily timestep. Grey area: simulated discharge from the different products

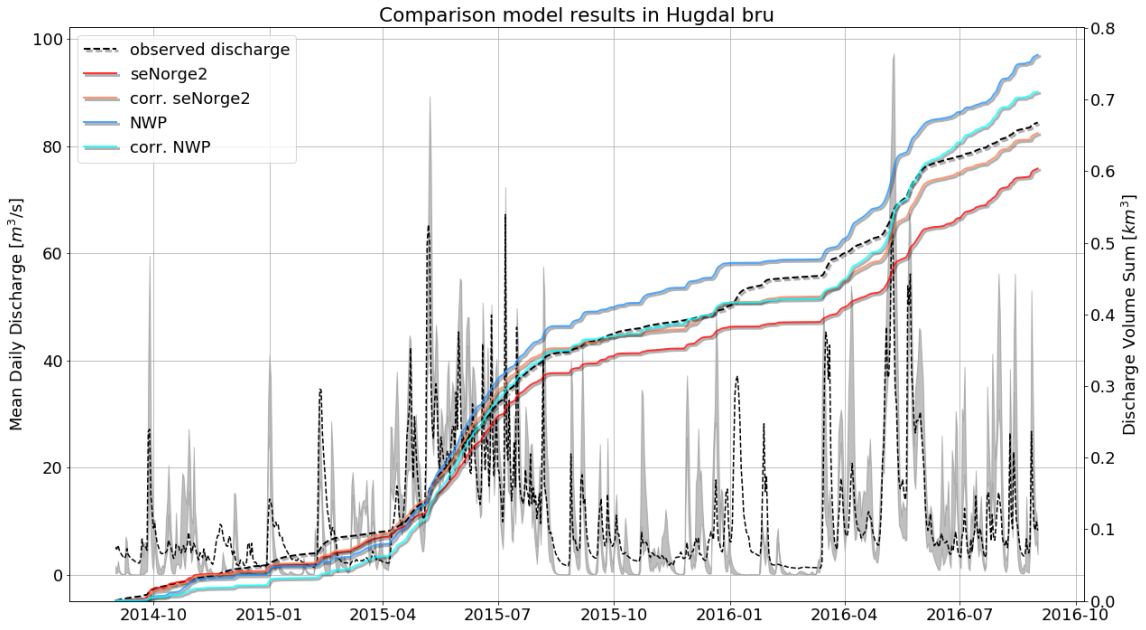


Figure 81: Comparison modelling results of Hugdal bru (Gaula catchment) with different datasets. The time series was averaged to a daily timestep. Grey area: simulated discharge from the different products

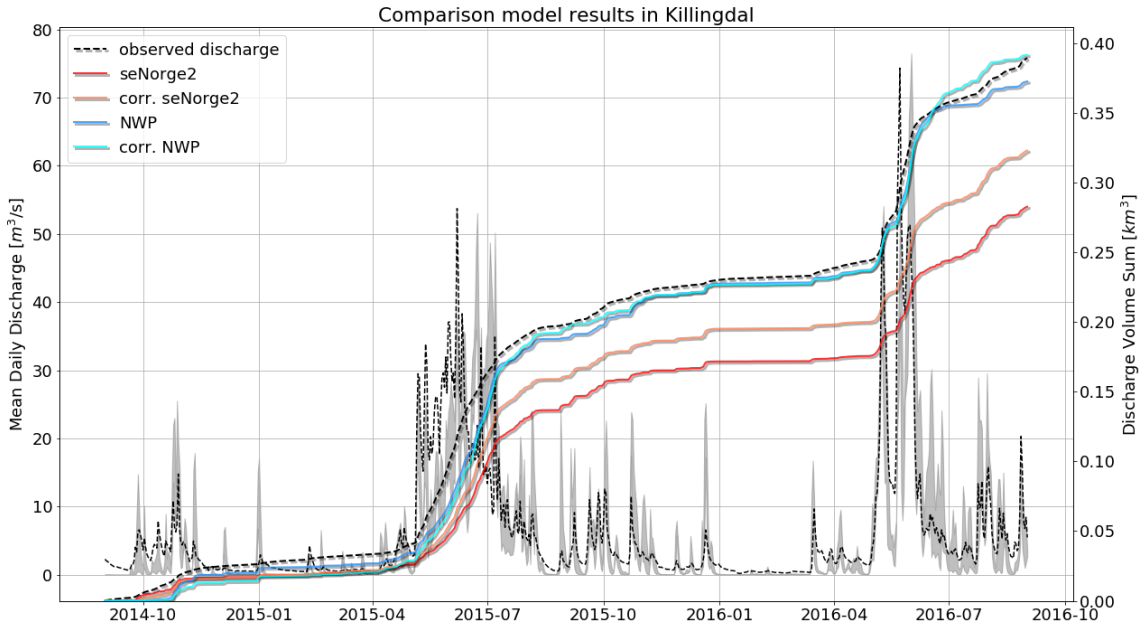


Figure 82: Comparison modelling results of Killingdal (Gaula catchment) with different datasets. The time series was averaged to a daily timestep. Grey area: simulated discharge from the different products

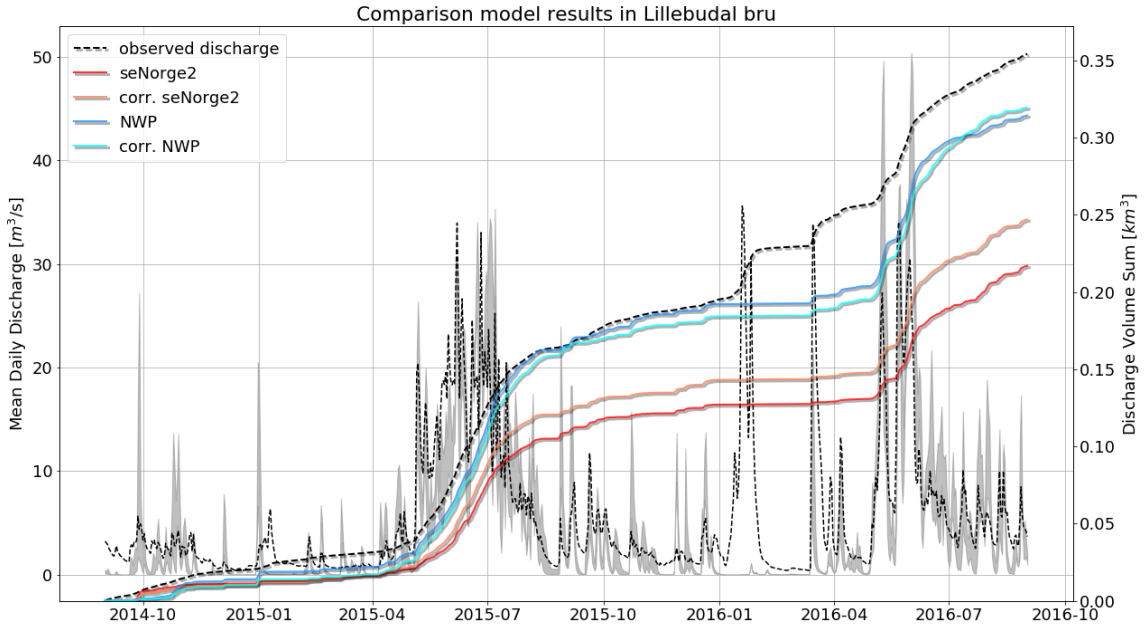


Figure 83: Comparison modelling results of Lillebudal bru (Gaula catchment) with different datasets. The time series was averaged to a daily timestep. Grey area: simulated discharge from the different products

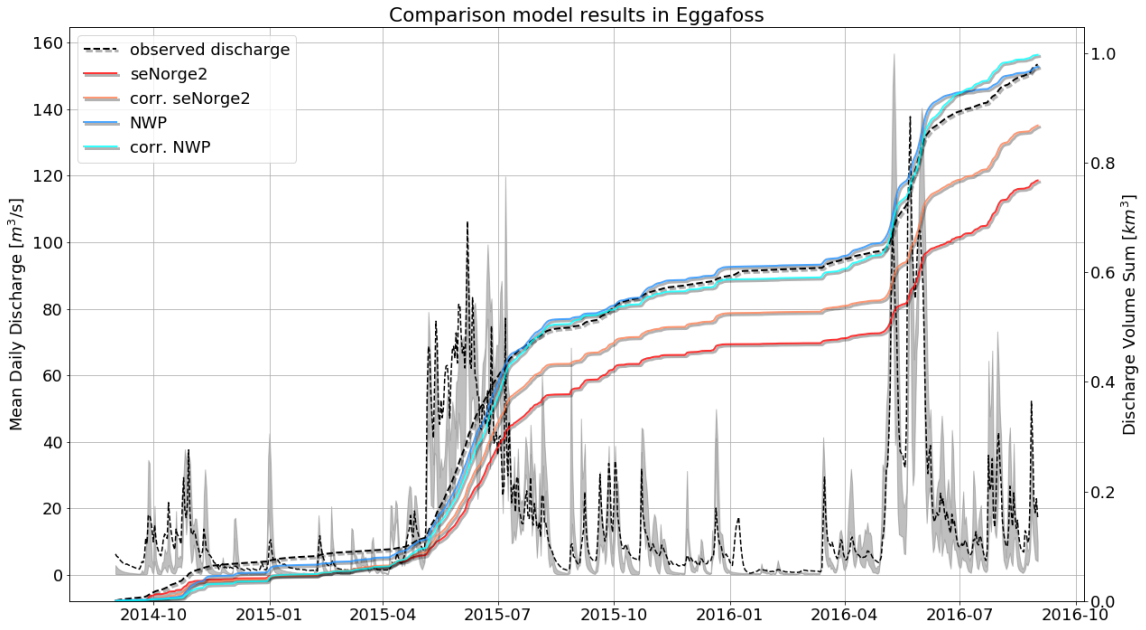


Figure 84: Comparison modelling results of Eggafoss (Gaula catchment) with different datasets. The time series was averaged to a daily timestep. Grey area: simulated discharge from the different products

F.1 Water balances

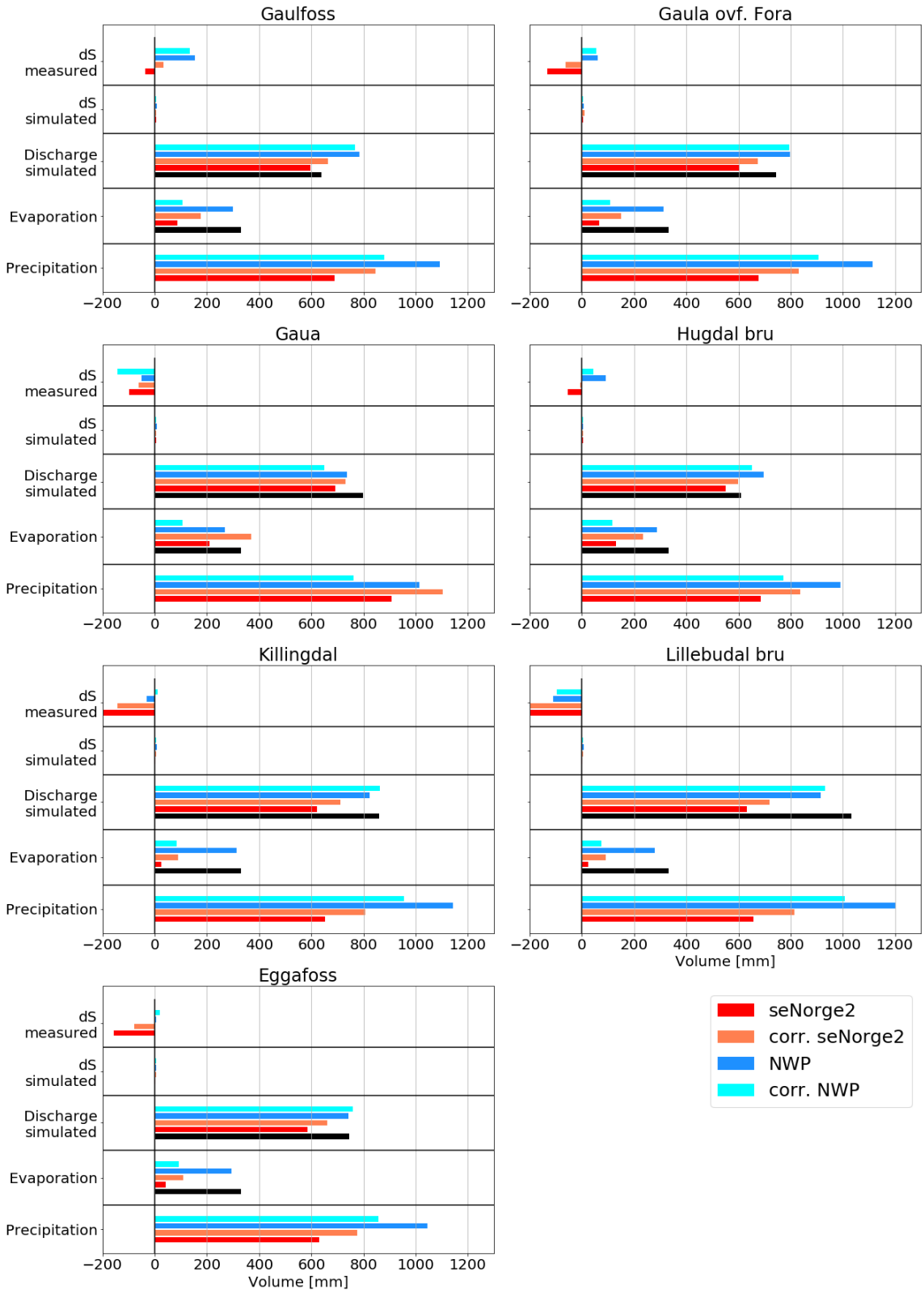


Figure 85: Annual water balance of Gaula

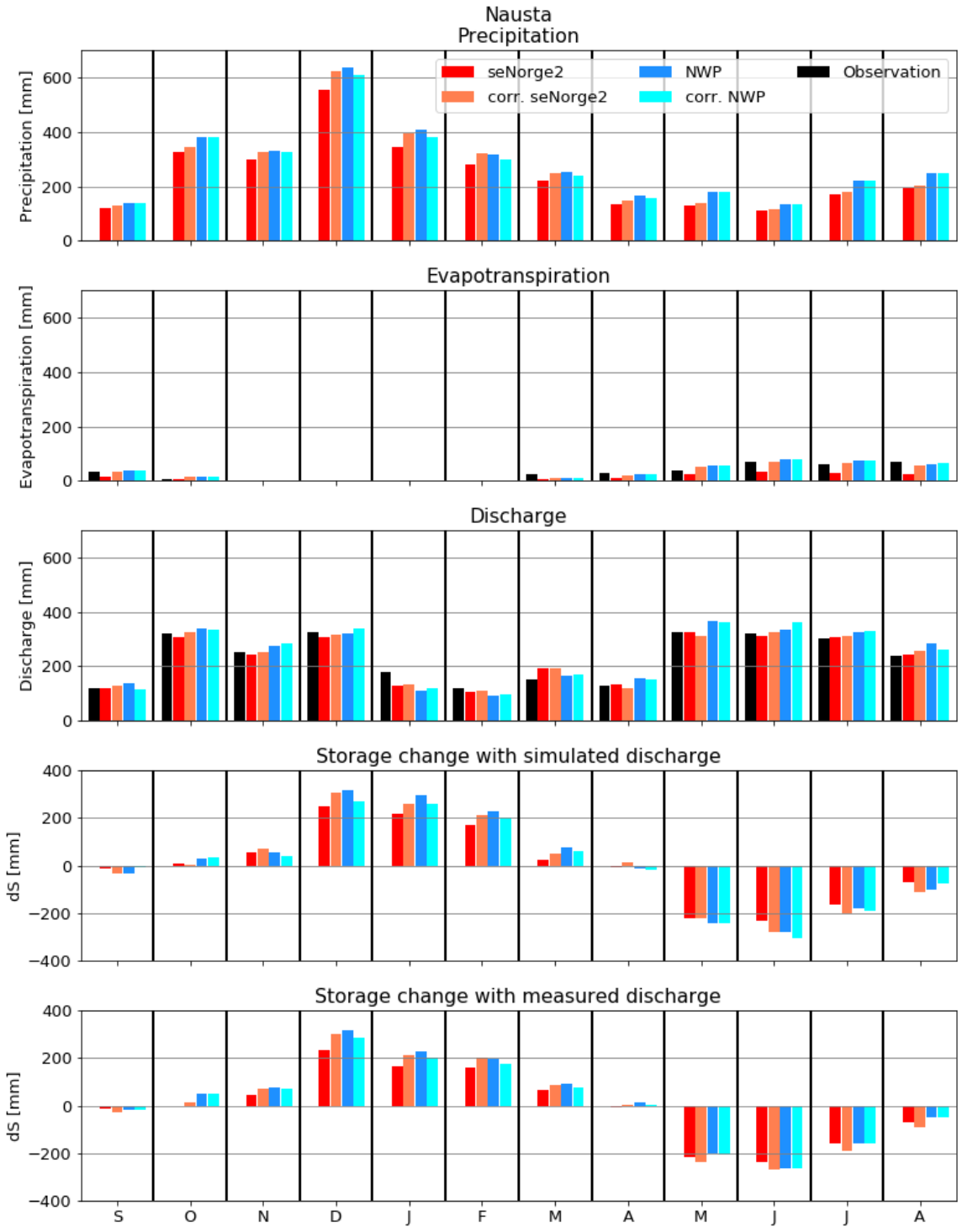


Figure 86: Seasonal water balance of Nausta when using different input products. Observed values are from discharge measurements and MOD16.

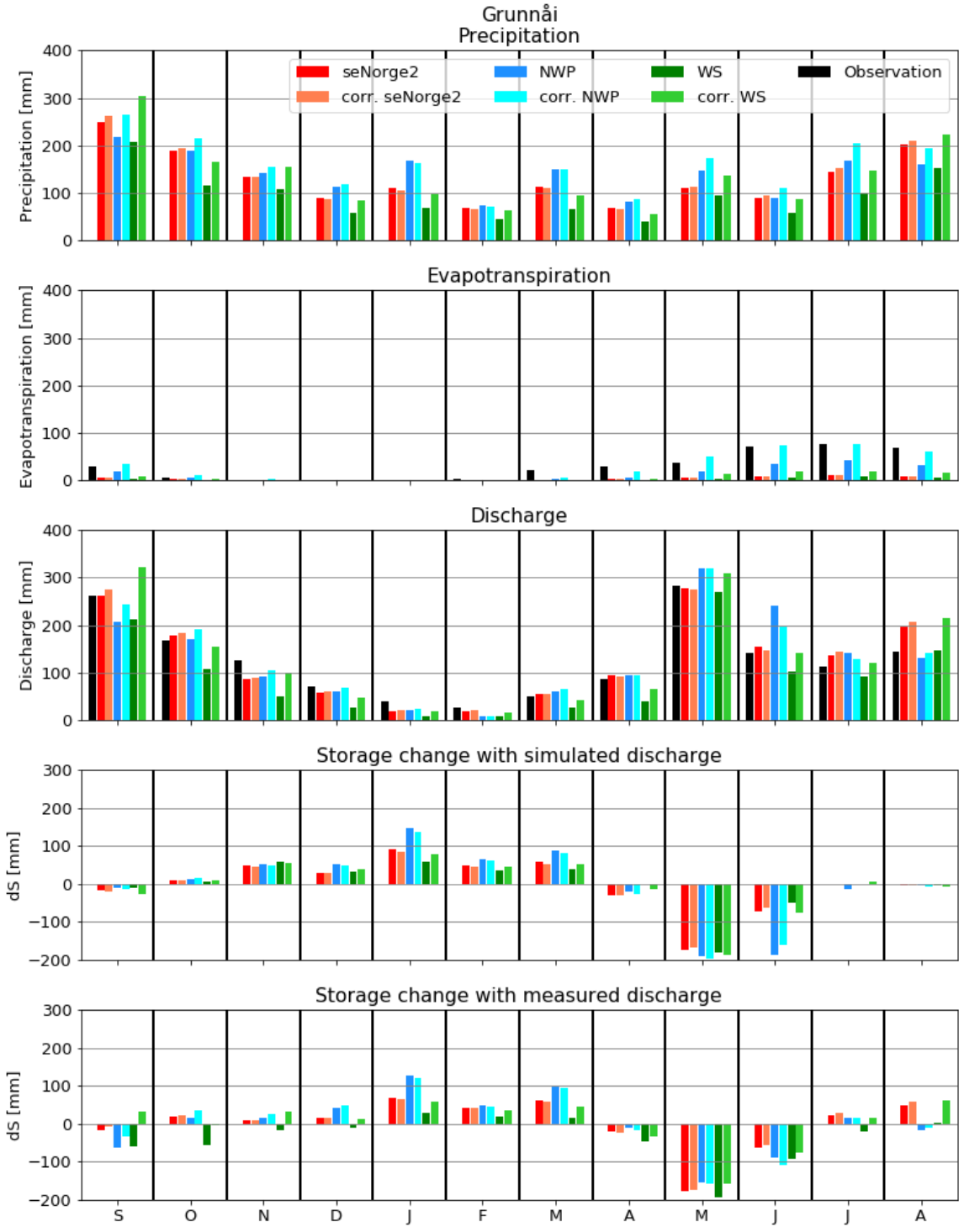


Figure 87: Seasonal water balance of Grunnåi when using different input products. Observed values are from discharge measurements and MOD16.

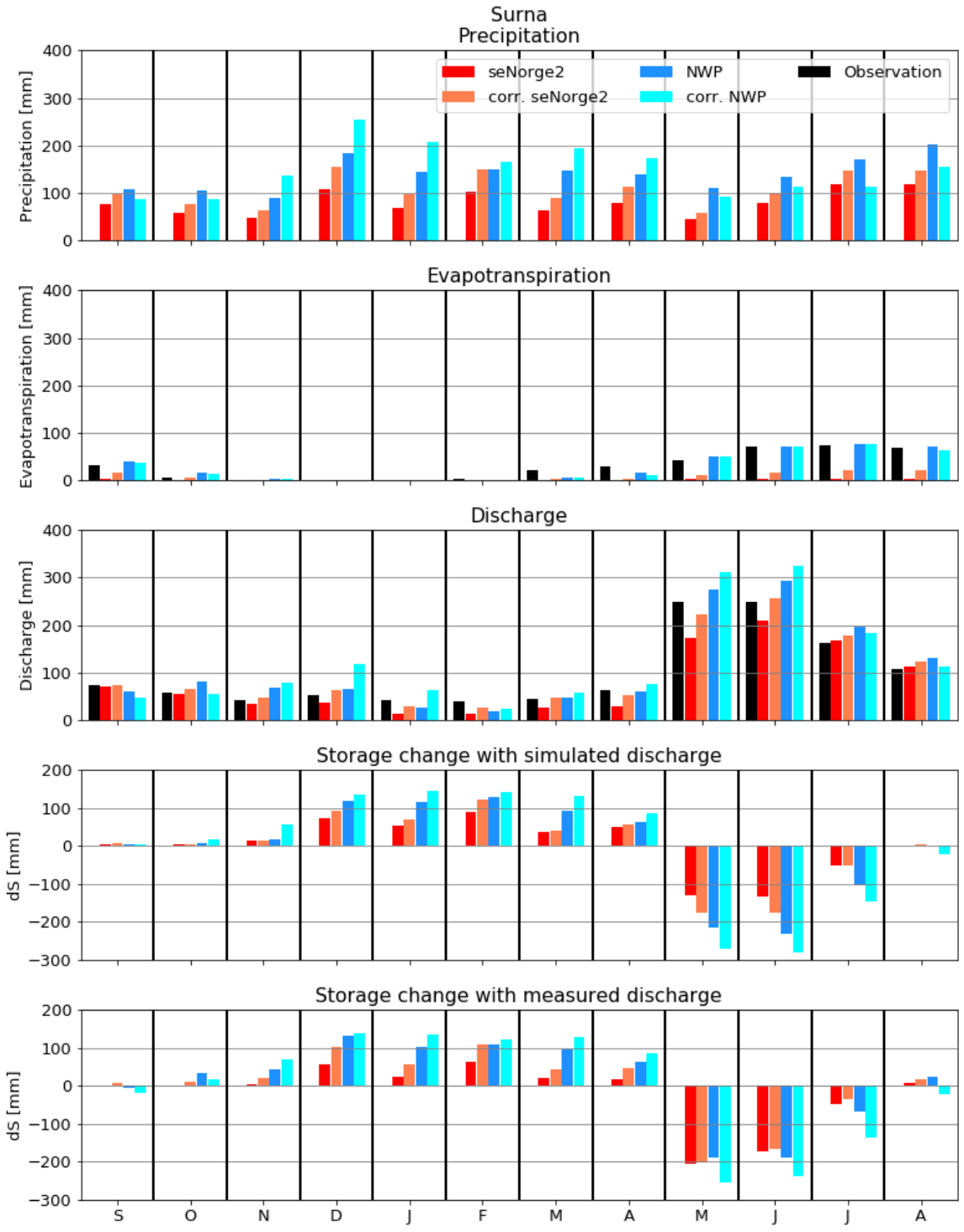


Figure 88: Seasonal water balance of Surna when using different input products. Observed values are from discharge measurements and MOD16.

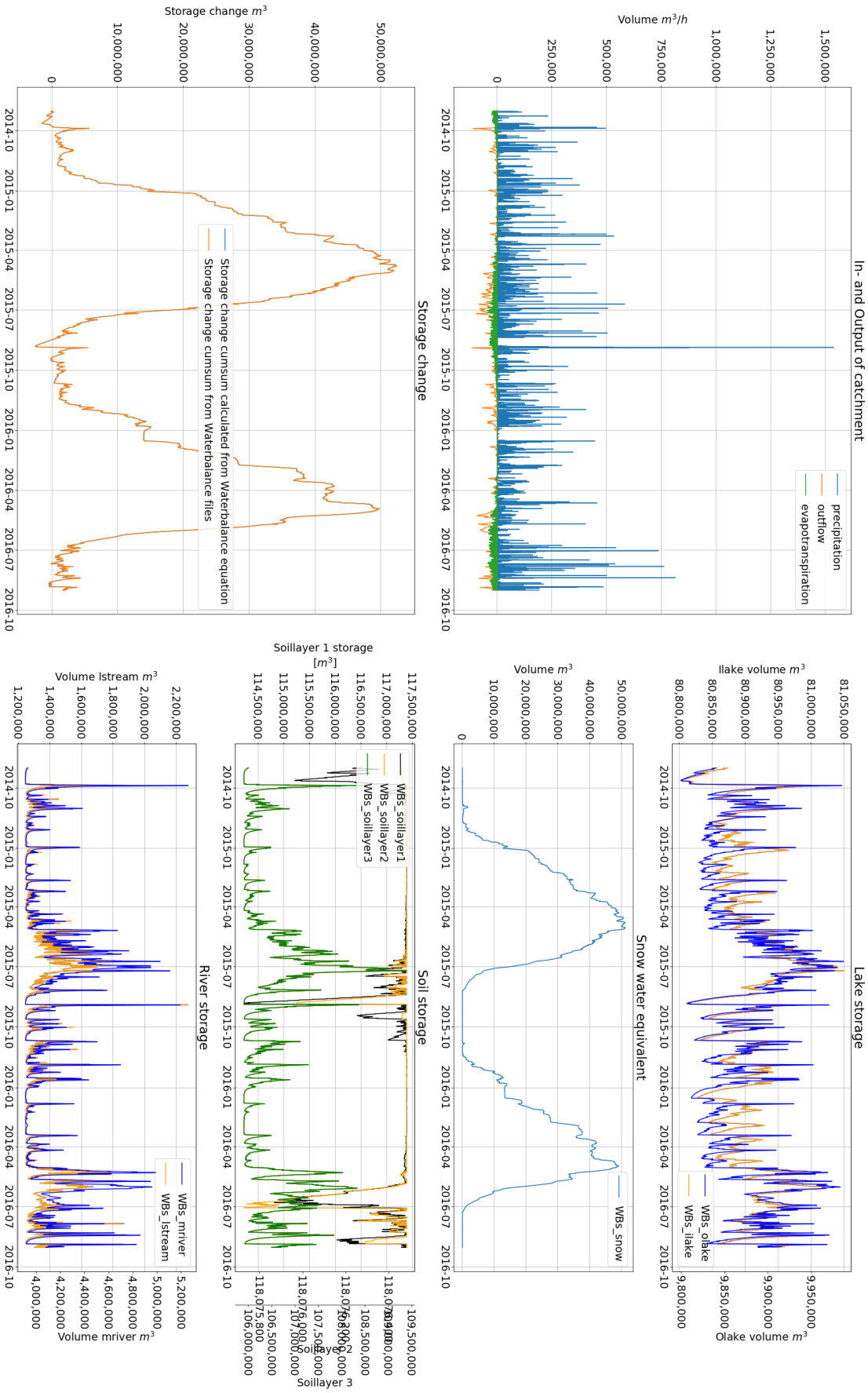


Figure 89: Example of the storage behaviour of the model within a catchment. Here: no storage accumulation. Catchment: Rinna, Dataset: NWP (uncorrected).

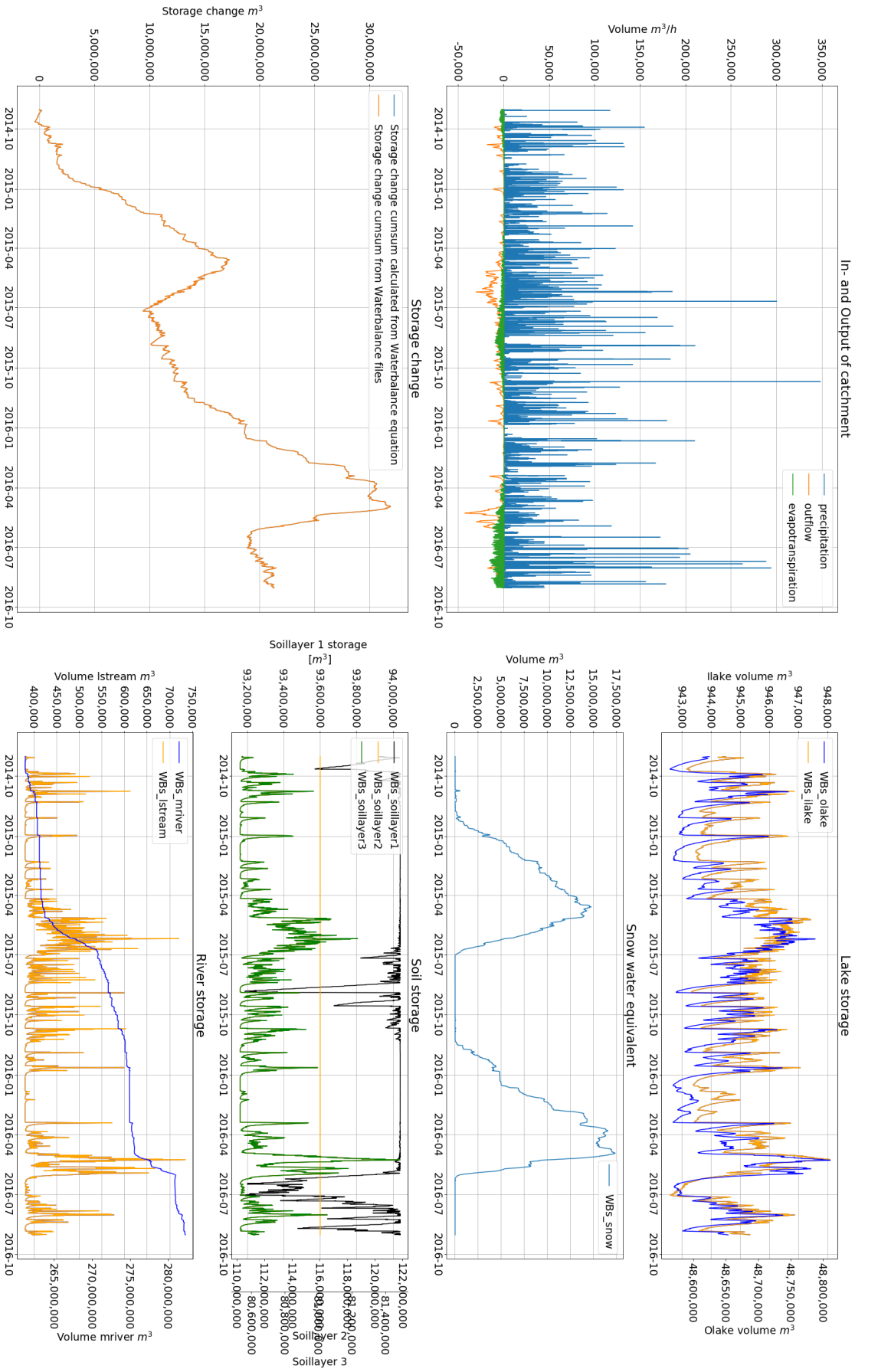


Figure 90: Example of the storage behaviour of the model within a catchment. Here: storage accumulation in the main river. Catchment: Usma, Dataset: NWP (uncorrected).

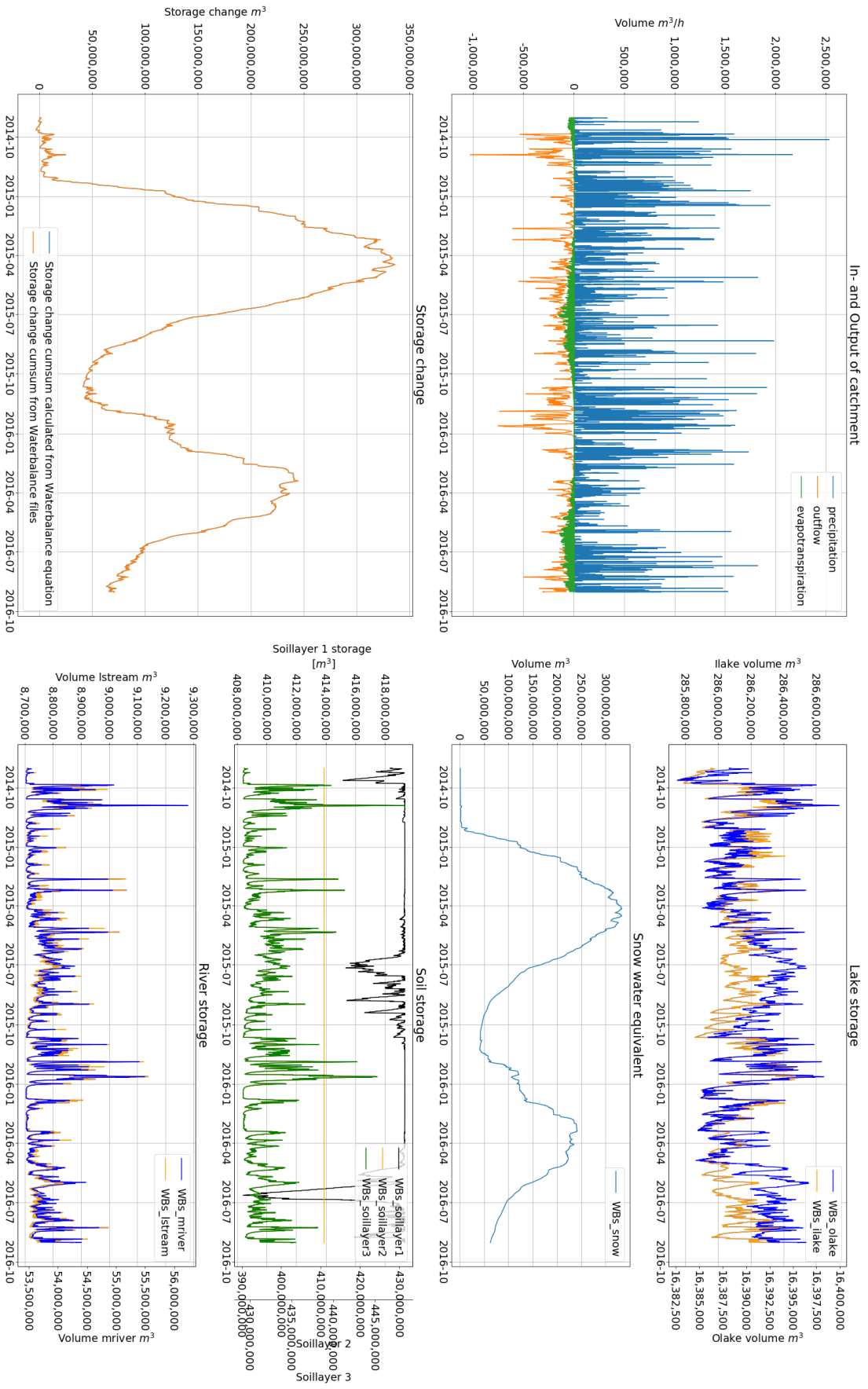


Figure 91: Example of the storage behaviour of the model within a catchment. Here: storage accumulation through snow. Catchment: Nausta, Dataset: NWP (uncorrected).

G Event Analysis

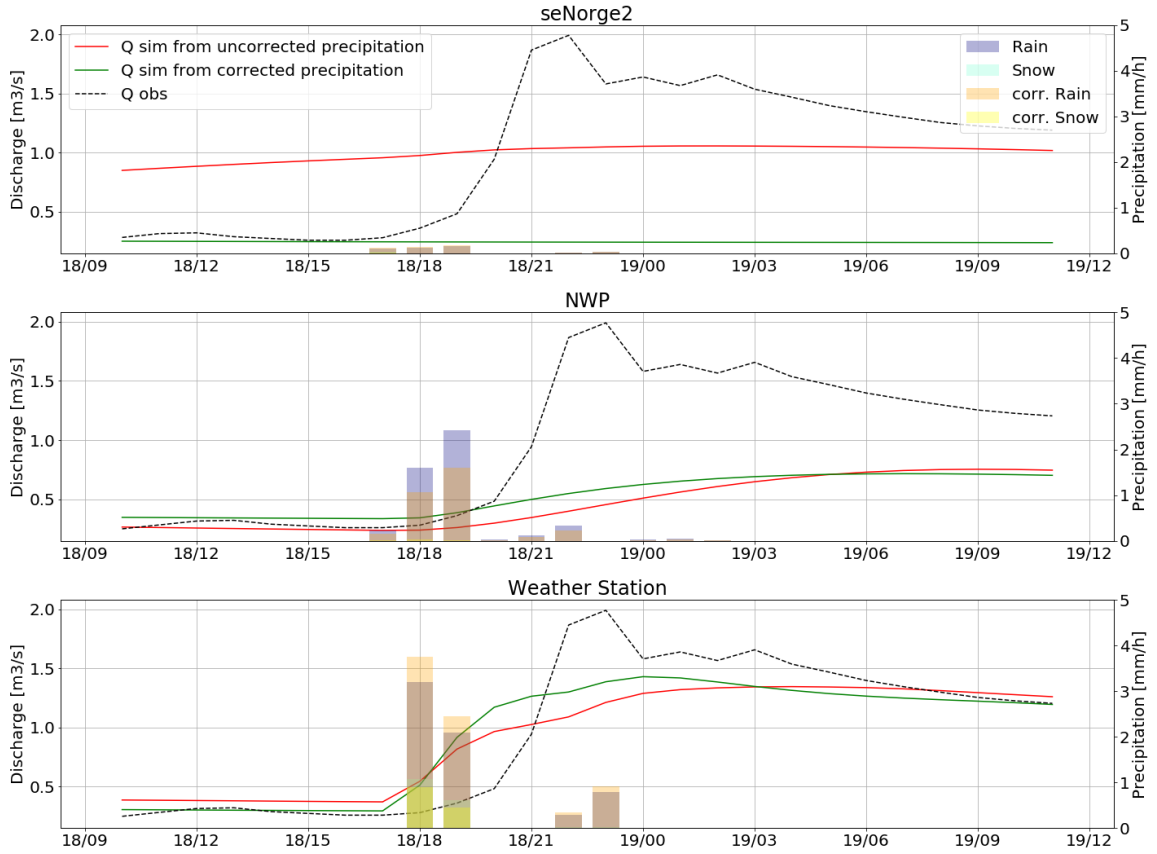


Figure 92: Precipitation event in Usma, 2014-10-18 to 2014-10-19. Figure includes observed and simulated discharge from weather station data, seNorge2 and NWP.

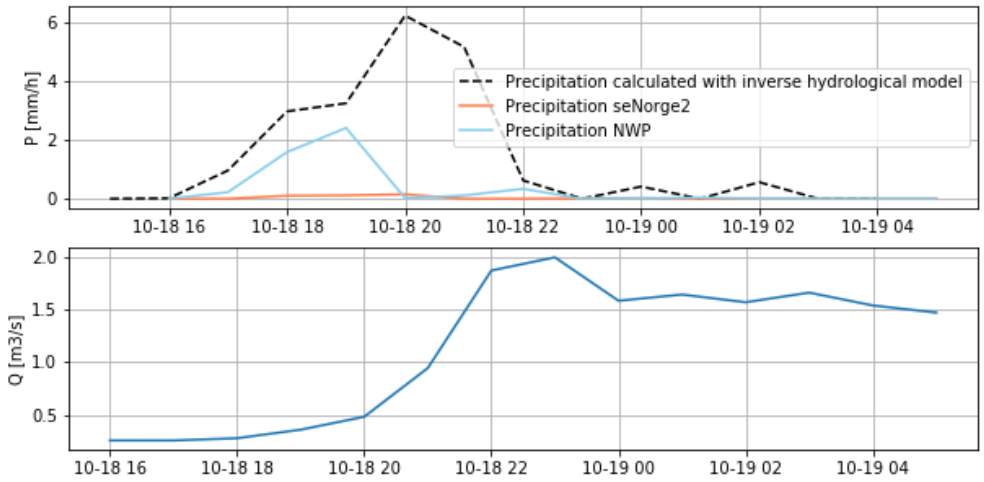


Figure 93: Precipitation estimated from seNorge2, NWP and an inverse hydrological model in Usma, October 2014

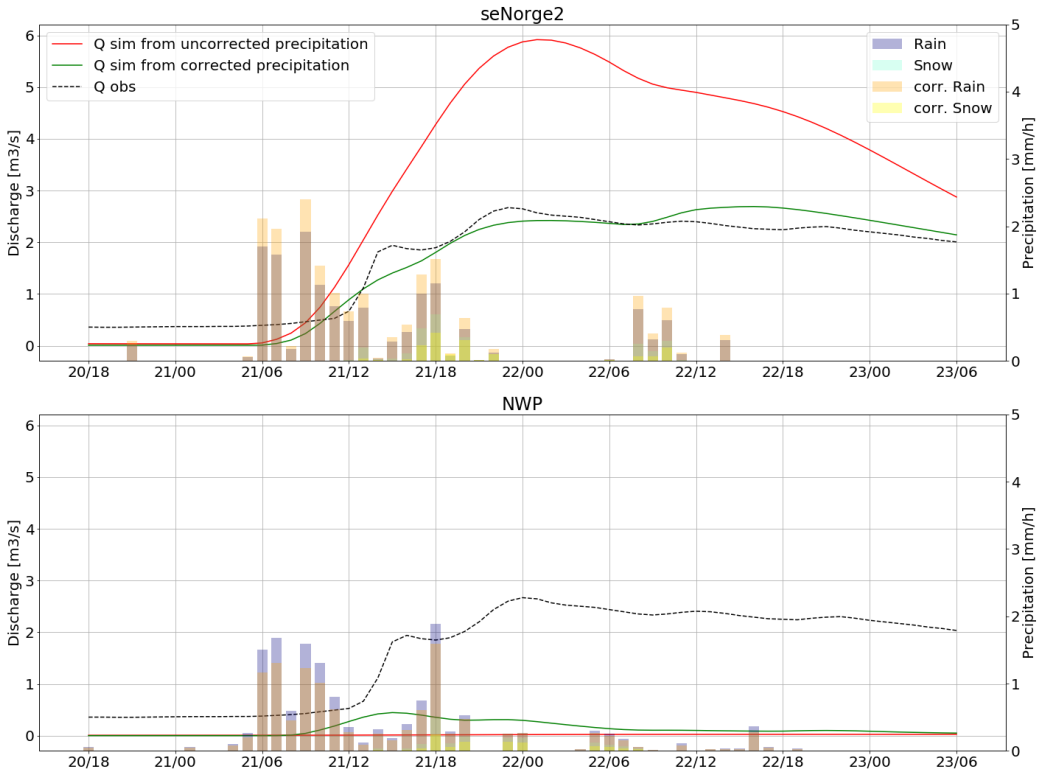


Figure 94: Precipitation Event in Rinna, 2014-09-21 to 2014-09-23. Figure includes observed and simulated discharge from, seNorge2 and NWP

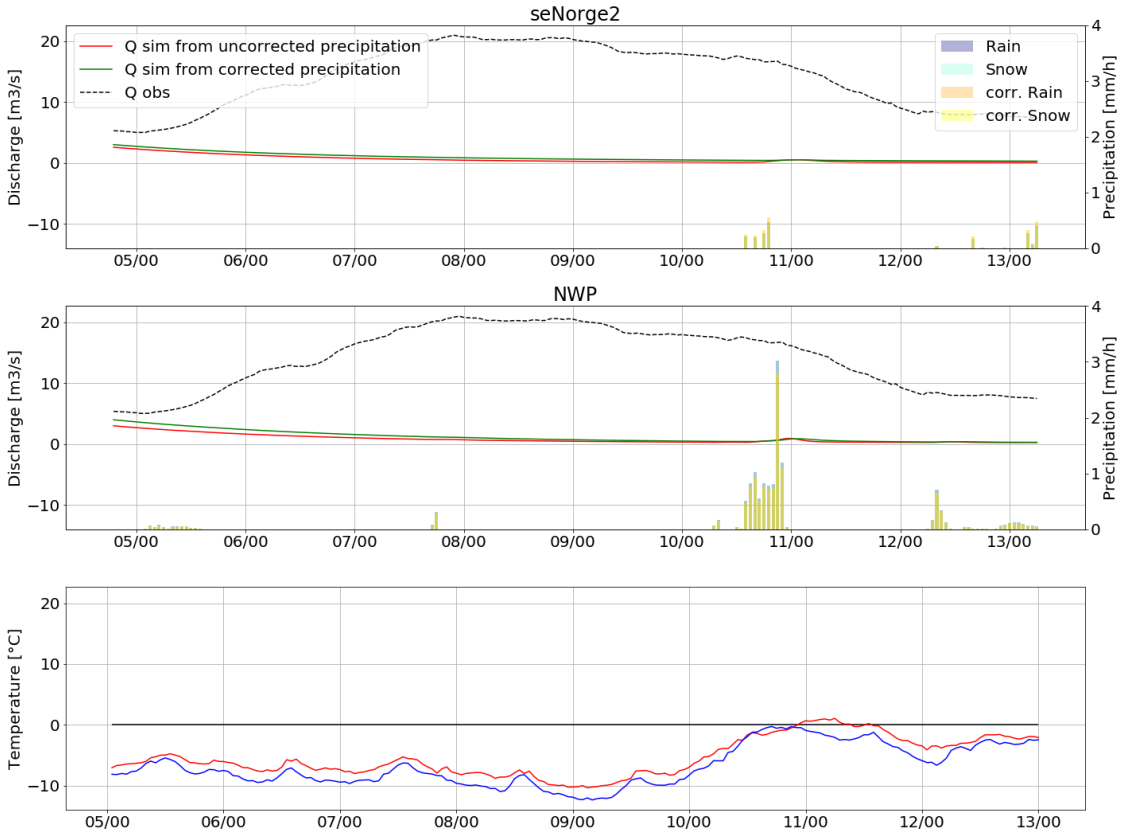


Figure 95: Precipitation Event in Nausta, 2016-01-05 to 2016-01-13. Figure includes observed and simulated discharge from, seNorge2 and NWP

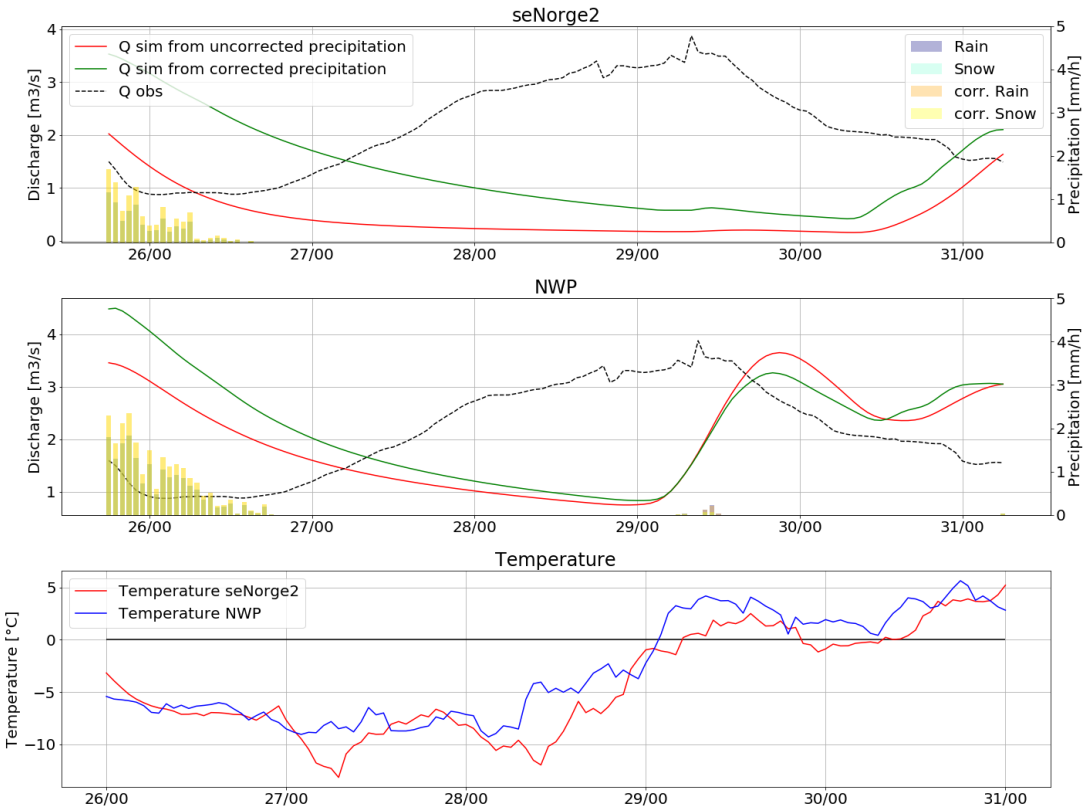


Figure 96: Precipitation Event in Rinna, 2015-12-26 to 2015-12-31. Figure includes observed and simulated discharge from, seNorge2 and NWP

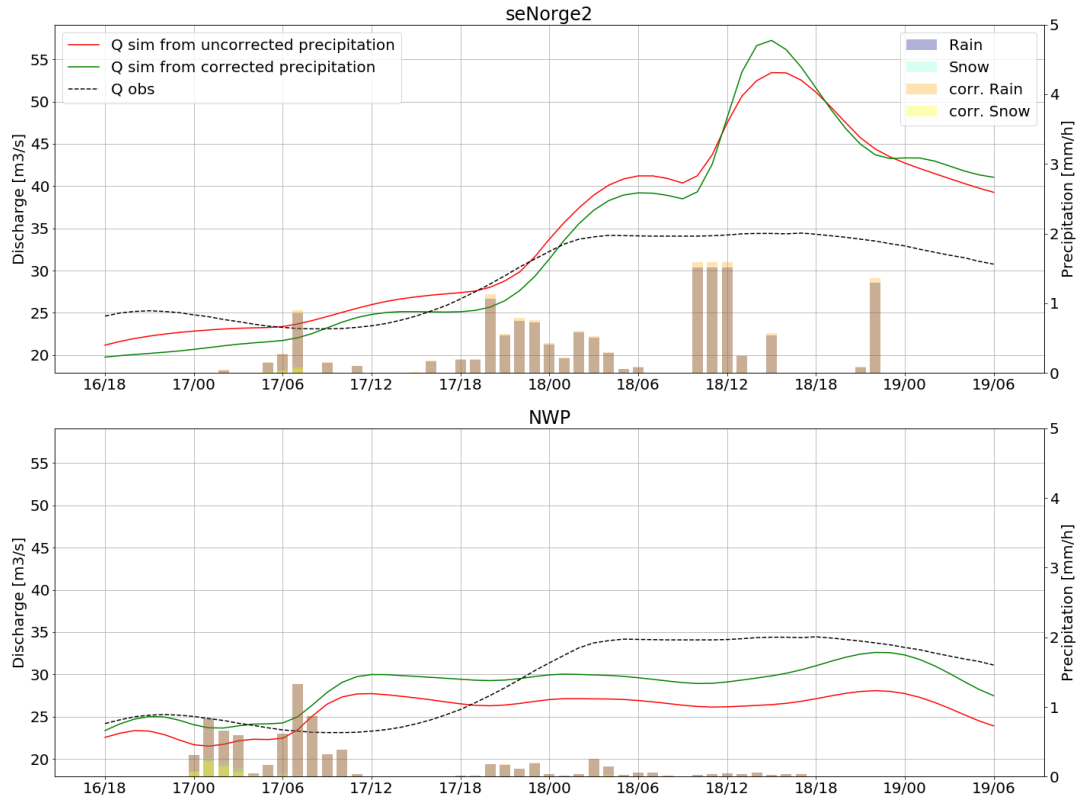


Figure 97: Precipitation Event in Grunnåi, 2016-06-16 to 2016-06-19. Figure includes observed and simulated discharge from, seNorge2 and NWP

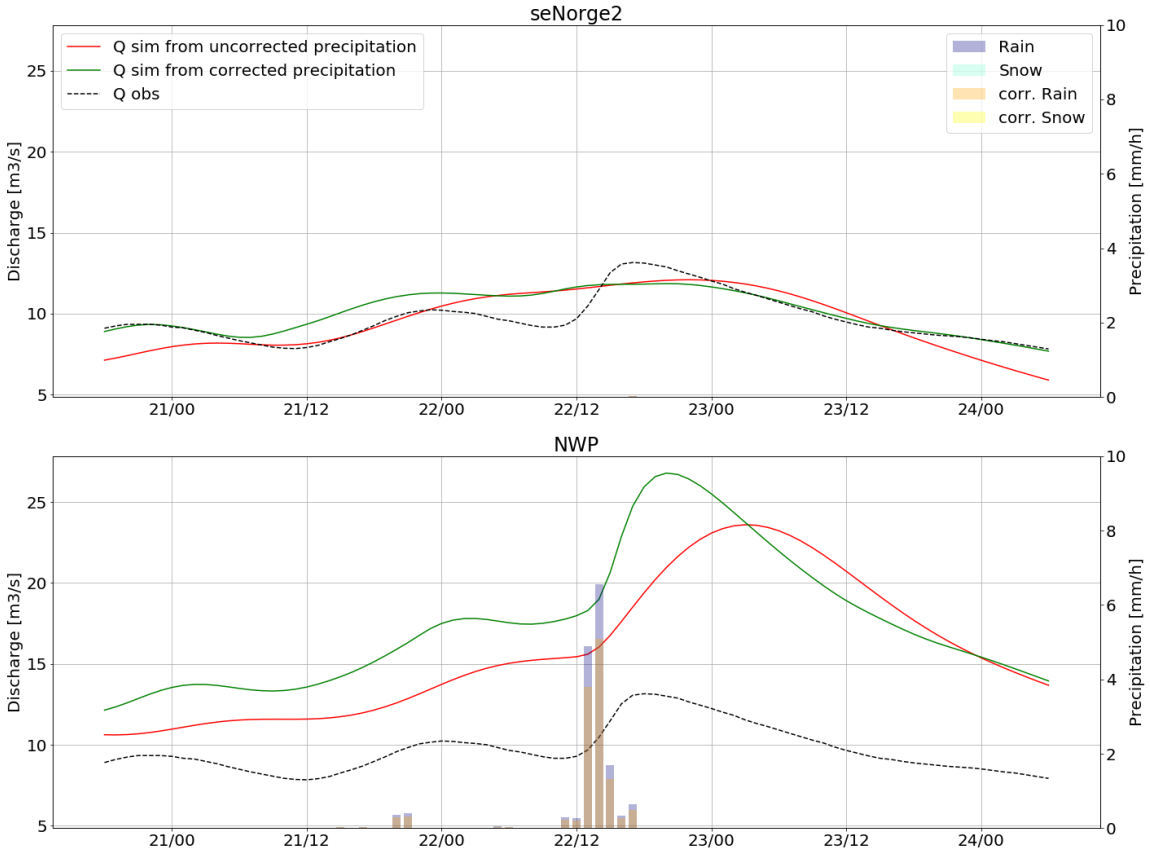


Figure 98: Precipitation Event in Rinna, 2015-06-21 to 2015-06-24. Figure includes observed and simulated discharge from, seNorge2 and NWP

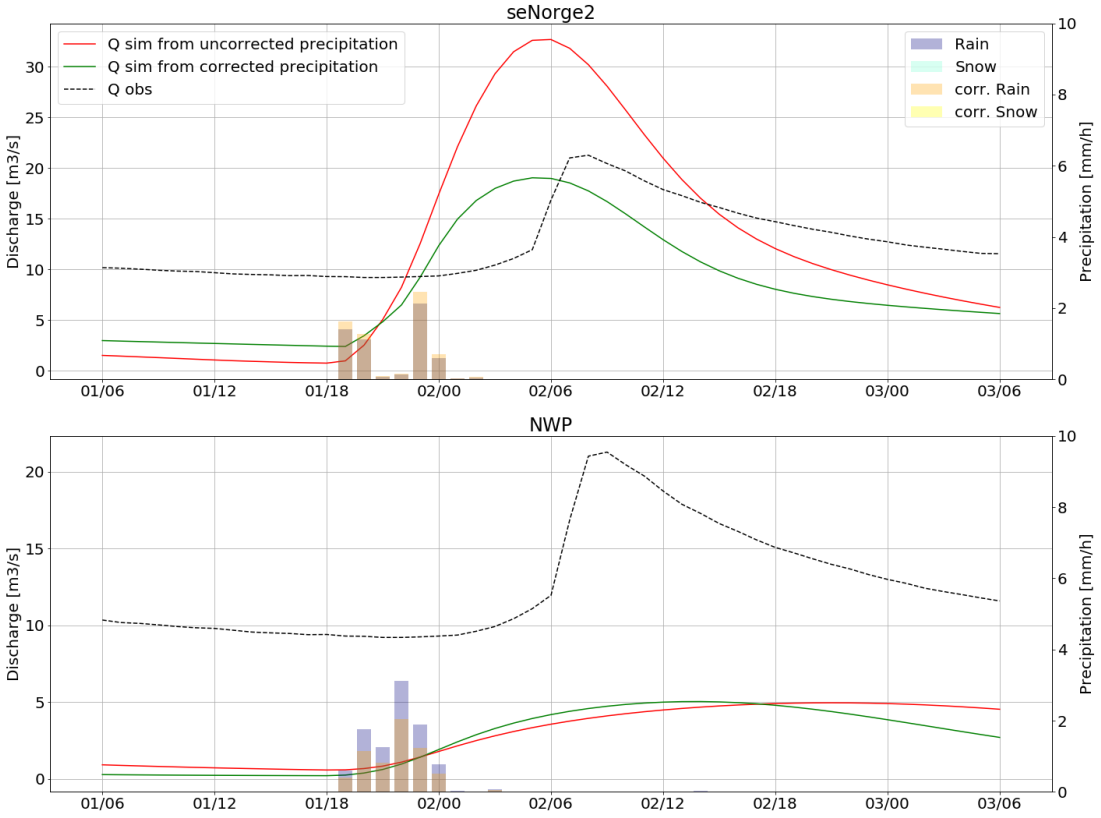


Figure 99: Precipitation Event in Eggafoss (Gaula), 2014-10-01 to 2014-10-03. Figure includes simulated and model discharge from, seNorge2 and NWP

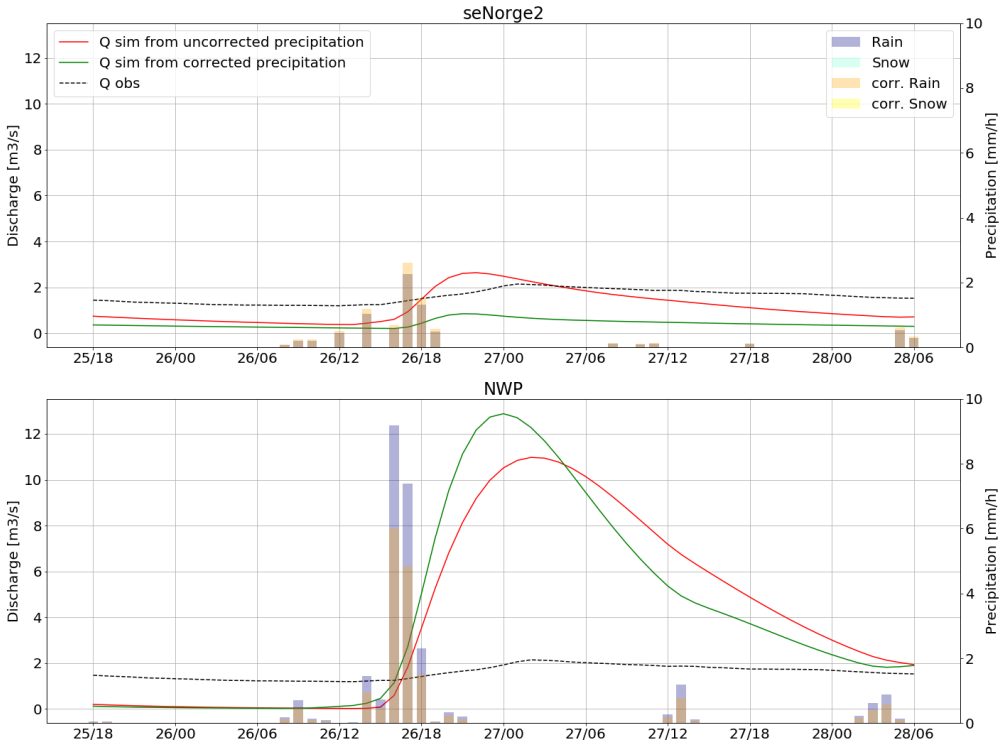


Figure 100: Precipitation Event in Gaua (Gaula=, 2016-07-26 to 2016-07-28. Figure includes simulated and model discharge from, seNorge2 and NWP

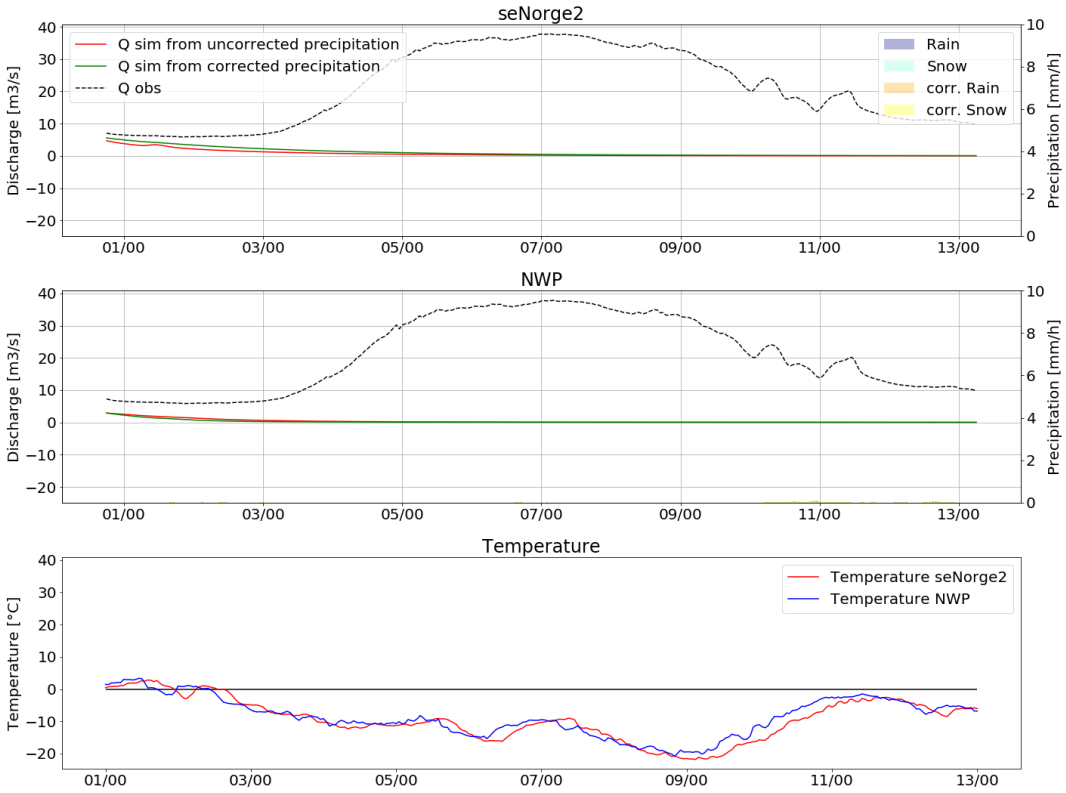
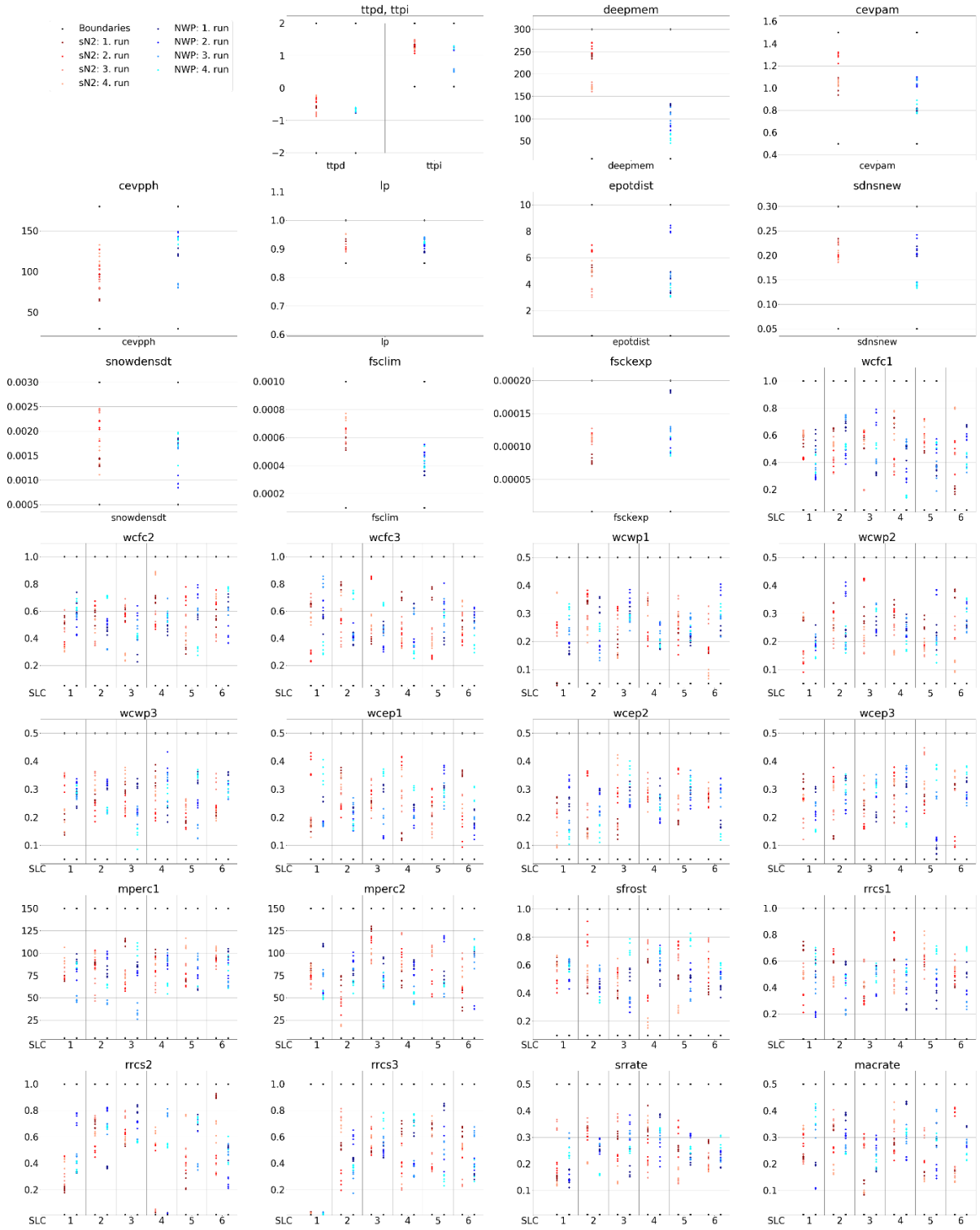


Figure 101: Precipitation Event in Hugdal bru, 2016-01-01 to 2016-01-13. Figure includes observed and simulated discharge from, seNorge2 and NWP

H Sensitivity Analysis

Figure 102: GLUE analysis in Grunnåi with normalized values.

Figure 103: GLUE analysis in Nausta with normalized values.


Boundaries
sN2: 1. run
sN2: 2. run
sN2: 3. run
sN2: 4. run
NWP: 1. run
NWP: 2. run
NWP: 3. run
NWP: 4. run
ttpd, ttpi
deepmem
cevpam
cevpph
lp
epotdist
sdnsnew
snowdensdt
fsclim
fsckexp
wcfc1
wcfc2
wcfc3
wcwp1
wcwp2
wcwp3
wcep1
wcep2
wcep3
mperc1
mperc2
sfrost
rrcs1
rrcs2
rrcs3
srrate
macrate
SLC

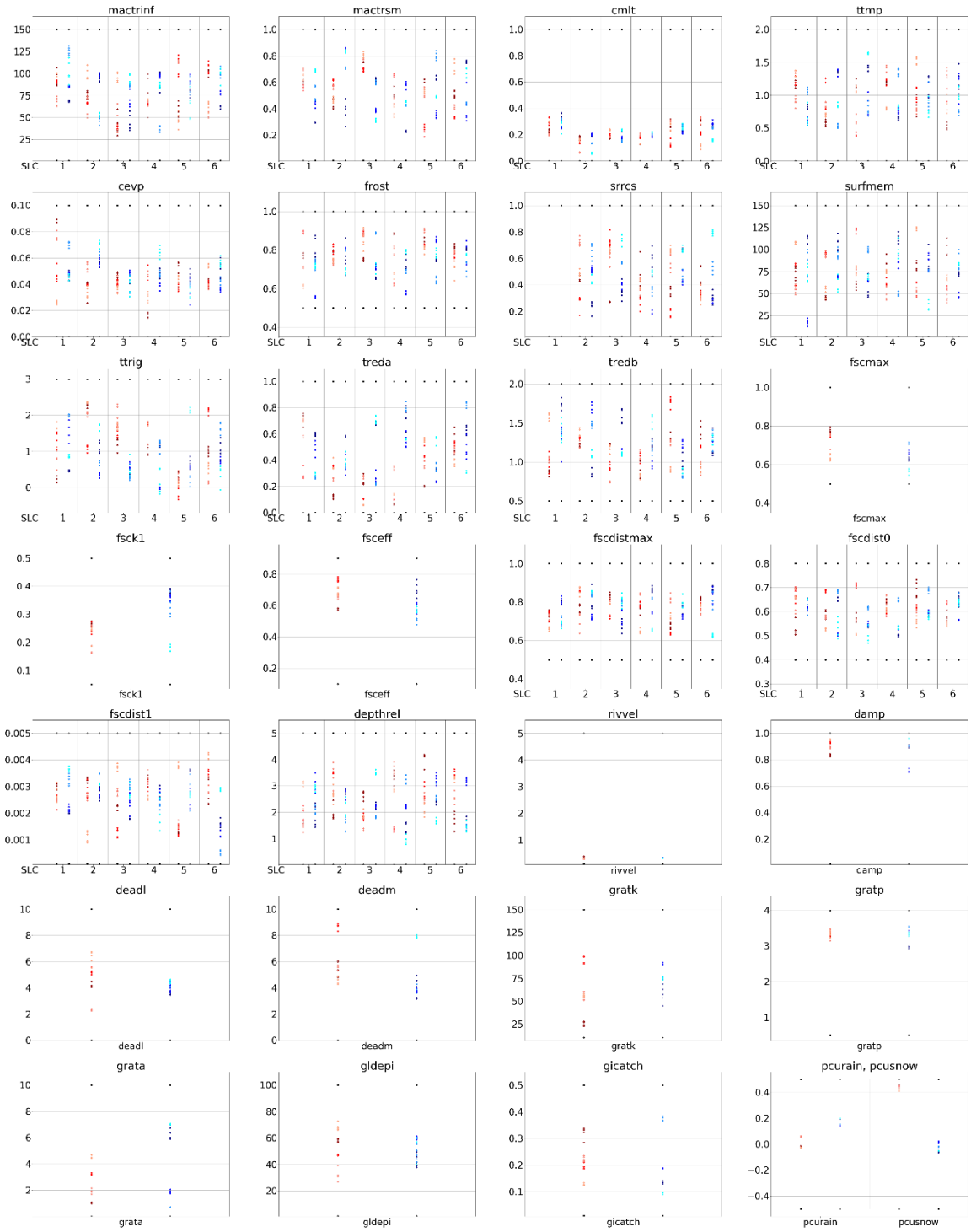
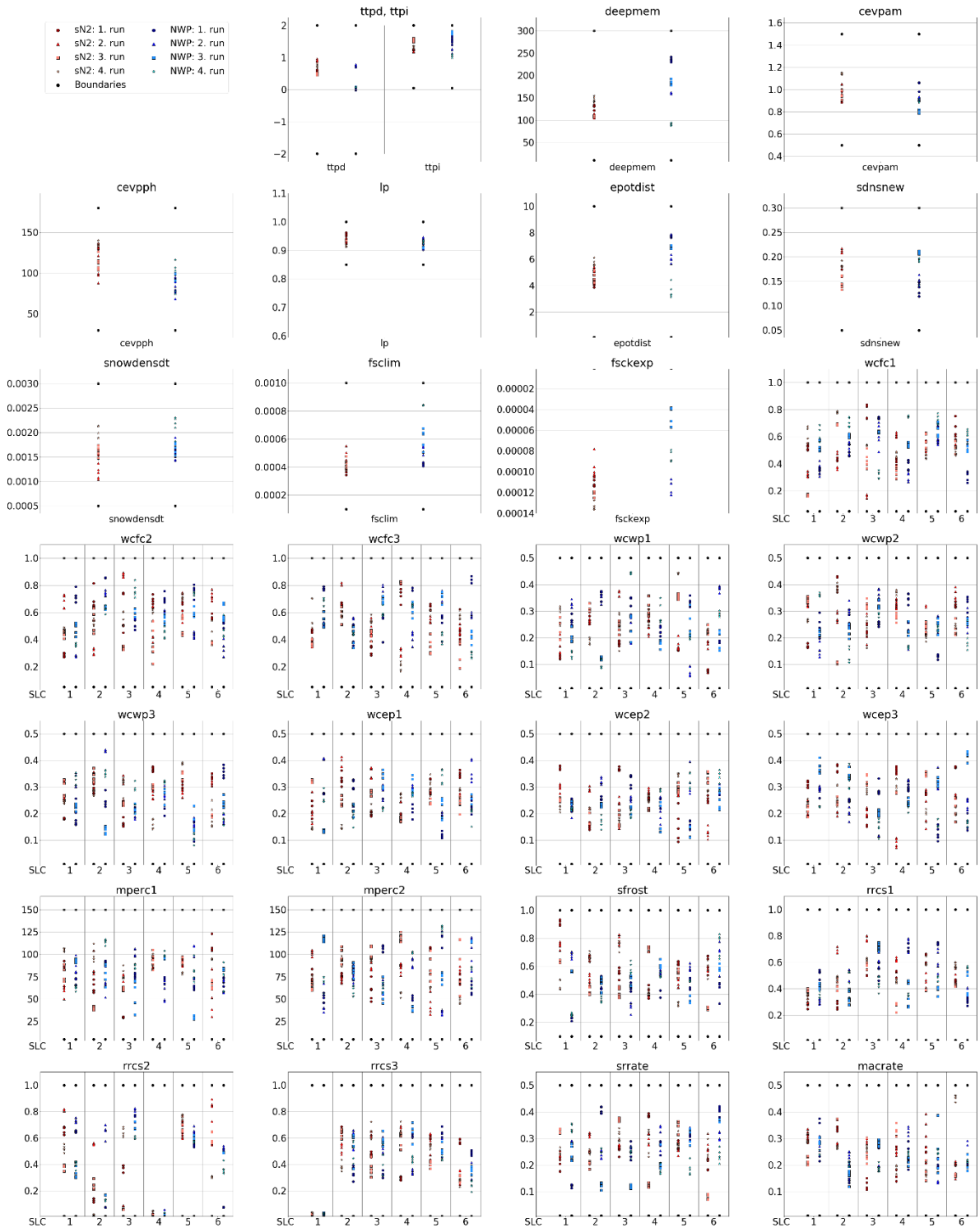


Figure 104: Parameter space of GLUE analysis in detail of Grunnåi.


sN2: 1. run
sN2: 2. run
sN2: 3. run
sN2: 4. run
Boundaries
NWP: 1. run
NWP: 2. run
NWP: 3. run
NWP: 4. run
ttpd, ttpi
deepmem
cevpam
cevpph
lp
epotdist
sdnsnew
snowdensdt
fsclim
fsckexp
wcfc1
wcfc2
wcfc3
wcwp1
wcwp2
wcwp3
wcep1
wcep2
wcep3
mperc1
mperc2
sfrost
rrcs1
rrcs2
rrcs3
srrate
macrate
SLC

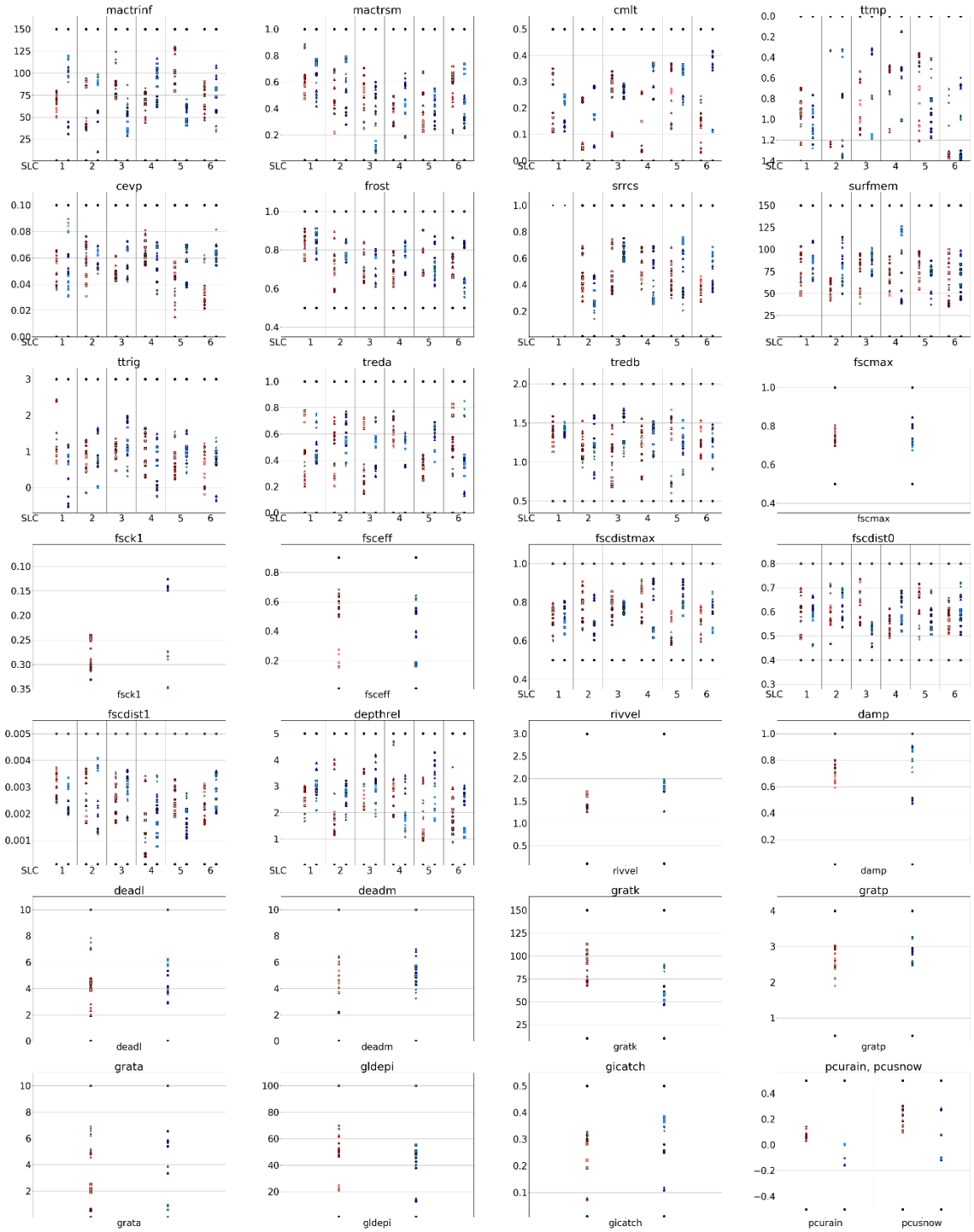


Figure 105: Parameter space of GLUE analysis of Nausta.

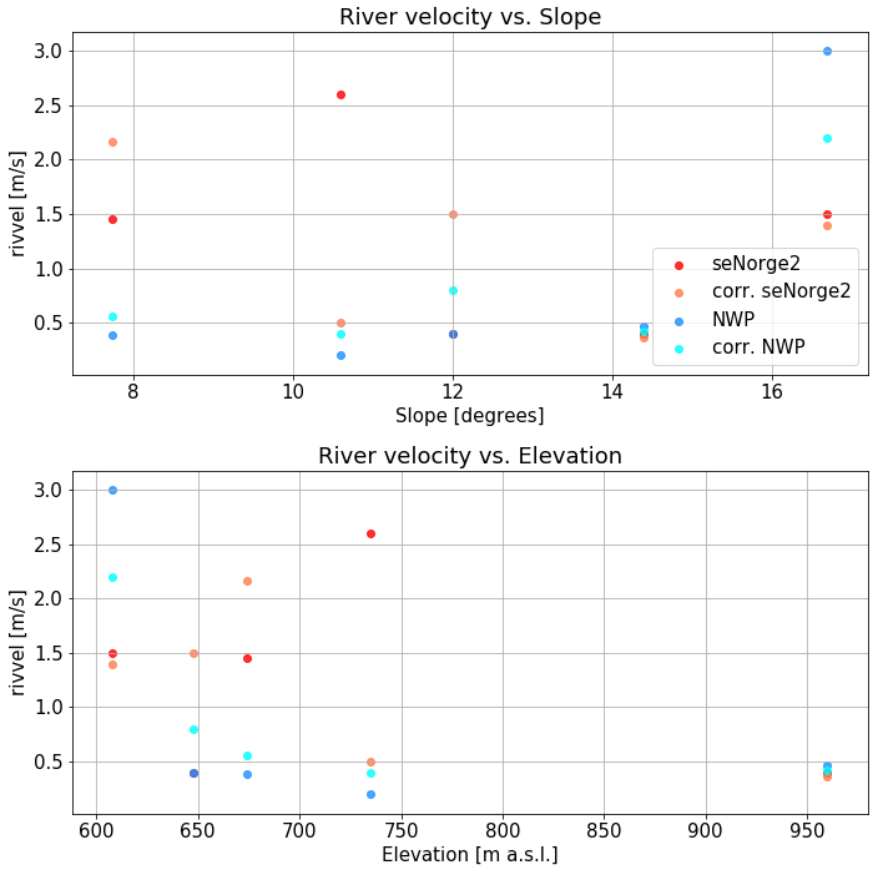


Figure 106: Relationship of calibrated river velocity with catchment mean slope and mean elevation

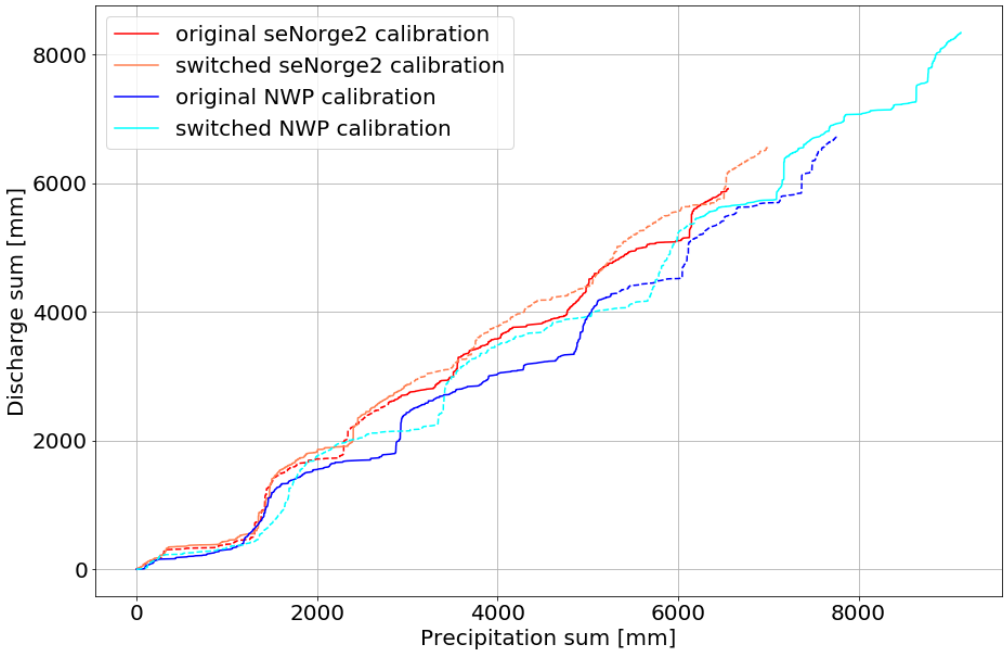


Figure 107: Double sum curve in Rinna, for seNorge2 and NWP and switched calibration and validation period. Dashed line: validation period

I Applied software and programmes in this thesis

In this thesis, different software for data manipulation, visualizing and sub-catchment adjustment was applied. Reading and extracting netcdf4 files was effectuated in RStudio (https://www.rstudio.com, version 5501.9.0.0). An integrated development environment for R programming. For handling and manipulating HYPE files a R package called HYPEtools (version 0.5.1) is available (https://github.com/rcapell/HYPEtools). This was used for spatial visualising HYPE results, as well as reading and interpolating `GeoData.txt` data.

To extract, adjust and recalculate the characteristics of sub-catchments WHIST (version 12.0.2, https://git.smhi.se/whist/whist/-/wikis/home) was used. This is a GIS-based tool from SMHI that is able to create and manipulate watersheds. The creation of watershed was done by Schönfelder (2017). In this thesis, WHIST was mainly used for small adjustments and recalculating `GeoData.txt` values. For further analysis and visualization QGIS (version 3.16.11, https://qgis.org/en/site) was applied.

Python 3.7.4 was applied for every other analysis. The packages subprocess (version 0.0.8, https://pypi.org/project/subprocess.run) and pandas (version 1.0.1, https://pypi.org/project/pandas/1.0.1) were the most necessary ones for handling and running HYPE from the console. The version of HYPE is 5.16.1 (https://sourceforge.net/projects/hype/files/).

J Digital appendix

The digital appendix is a summary of all important files that were created during this thesis. It contains models, figures, scripts and videos created during this master thesis. The appendix is structured as follows:

1. Manuscript and Presentation

Contains the digital version of this thesis as well as slides from the final presentation.

2. HYPE models

Contains all models of the catchments, divided into datasets. “ForcingData” contains observational data from the catchment. The folder ‘M0-0’ contains the initial calibration, ‘results1’ contains the model version after the stepwise calibration.

3. Figures

Contains all figures created during this master thesis. The folder structure is similar to the second folder with each catchment and dataset. Additionally, figures and videos from precipitation events, the GLUE analysis, MOD16 data and seasonal water balances are included.

3. Scripts

Contains the Python and R scripts developed in this study. The excel sheet “SummaryScriptsJessica” summarizes and explains the scripts and their location.

X. Appendix

Contains the results from the low-flow calibration.

Publication bibliography

Abdella, Yisak Sultan; Engeland, Kolbjoern; Lepioufle, Jean-Marie (2012): Evaluation of radar-derived precipitation estimates using runoff simulation : report for the NFR Energy Norway funded project 'Utilisation of weather radar data in atmospheric and hydrological models'.

Acharya, Rajat (2017): Satellite Signal Propagation, Impairments and Mitigation: Academic Press. ISBN: 978-0-12-809732-8.

Al-jabery, Khalid K.; Obafemi-Ajayi Tayo; Olbricht, Gayla R.; Wunsch II, Donald C.; Obafemi-Ajayi, Tayo: Evaluation of cluster validation metrics / Evaluation of cluster validation metrics. In *Khalid K. Al-jabery, Obafemi-Ajayi Tayo, Gayla R. ,Olbricht, Donald C. Wunsch II (Eds.): Computational Learning Approaches to Data Analytics in Biomedical Applications: Academic Press*, pp. 189–208. DOI: 10.1016/B978-0-12-814482-4.00007-3.

Allan, Richard; Pereira, L.; Smith, Martin (1998): Crop evapotranspiration-Guidelines for computing crop water requirements. In *FAO irrigation and drainage paper 56*.

Anwar, Faizan (2015): An Analysis of MODIS Evapotranspiration Products in Peru: Analysis of Spatial and Temporal Patterns. Independent Study. University of Stuttgart, Germany. Water Resources Engineering and Management.

Arheimer, B.; Lindström, G. (2013): Implementing the EU Water Framework Directive in Sweden. In *G. Blöschl, M. Sivapalan, T. Wagener, A. Viglione, & H. Savenije (Eds.), Runoff Prediction in Ungauged Basins: Synthesis across Processes, Places and Scales. Cambridge: Cambridge University Press*. DOI: 10.1017/CBO9781139235761.014.

Arheimer, B.; Pimentel, R.; Isberg, K.; Crochemore, L.; Andersson, J. C. M.; Hasan, A.; Pineda, L. (2020): Global catchment modelling using World-Wide HYPE (WWH), open data, and stepwise parameter estimation. In *Hydrol. Earth Syst. Sci.* 24 (2), pp. 535–559. DOI: 10.5194/hess-24-535-2020.

Auer, Hannah (2018): Evaluation of ensemble precipitation forecast of selected catchments in Rhineland-Palatinate. Master Thesis. Universität Stuttgart. IWS.

Bahr, David B.; Meier, Mark F.; Peckham, Scott D. (1997): The physical basis of glacier volume-area scaling. In *Journal of geophysical research* 102, 20355–20362. DOI: 10.1029/97JB01696.

Bahr, David B.; Pfeffer, W. Tad; Kaser, Georg (2015): A review of volume-area scaling of glaciers. In *Reviews of Geophysics* 53 (1), pp. 95–140. DOI: 10.1002/2014RG000470.

Bárdossy, András; Kilsby, Chris; Birkinshaw, Stephen; Wang, Ning; Anwar, Faizan (2022): Is Precipitation Responsible for the Most Hydrological Model Uncertainty? In *Front: Water (Frontiers in Water)* 4. DOI: 10.3389/frwa.2022.836554.

Beck, H. E.; van Dijk, A. I. J. M.; Roo, A. de; Dutra, E.; Fink, G., Orth, R.; Schellekens, J. (2017): Global evaluation of runoff from 10 state-of-the-art hydrological models. In *Hydrology and Earth System Sciences* 21, pp. 2881–2903. DOI: 10.5194/hess-21-2881-2017.

Beldring, S.; Engeland, K.; Roald, L. A.; Sælthun, N. R.; Voksø, A. (2003): Estimation of parameters in a distributed precipitation-runoff model for Norway. In *Hydrol. Earth Syst. Sci.* 7 (3), pp. 304–316. DOI: 10.5194/hess-7-304-2003.

Berg, Peter; Norin, Lars; Olsson, Jonas (2016): Creation of a high resolution precipitation data set by merging gridded gauge data and radar observations for Sweden. In *Journal of Hydrology* 541 A, pp. 6–13. DOI: 10.1016/j.jhydrol.2015.11.031.

Berne, A.; Krajewski, W. (2013): Radar for hydrology: Unfulfilled promise or unrecognized potential? In *Advances in Water Resources* 51, pp. 357–366. DOI: 10.1016/j.advwatres.2012.05.005.

Berne, Alexis; Delrieu, Guy; Creutun, Jean; Obled, Charles (2004): Temporal and Spatial Resolution of Rainfall Measurements Required for Urban Hydrology. In *Journal of Hydrology* 299 (3-4), pp. 166–179. DOI: 10.1016/j.jhydrol.2004.08.002.

Beven, Keith; Binley, Andrew (1992): The future of distributed models: Model calibration and uncertainty prediction. In *Hydrol. Process.* 6 (3), pp. 279–298. DOI: 10.1002/hyp.3360060305.

Beven, Keith; Binley, Andrew (2014): GLUE: 20 years on. In *Hydrol. Process.* 28 (24), pp. 5897–5918. DOI: 10.1002/hyp.10082.

Borup, Morten; Grum, Morten; Linde, Jens Jørgen; Mikkelsen, Peter Steen (2016): Dynamic gauge adjustment of high-resolution X-band radar data for convective rain storms: Model-based evaluation against measured combined sewer overflow. In *Journal of Hydrology* 539, pp. 687–699. DOI: 10.1016/j.jhydrol.2016.05.002.

Boslaugh, Sarah; Watters, Paul Andrew (2008): Statistics in a Nutshell: A Desktop Quick Reference: O'Reilly Media. ISBN: 978-0596510497.

Boudevillain, Brice; Delrieu, Guy; Wijbrans, Annette; Confoland, Audrey (2016): A high-resolution rainfall re-analysis based on radar–raingauge merging in the Cévennes-Vivarais region, France. In *Journal of Hydrology* 541, pp. 14–23. DOI: 10.1016/j.jhydrol.2016.03.058.

Braak, Cajo J. F. ter (2006): A Markov Chain Monte Carlo version of the genetic algorithm Differential Evolution: easy Bayesian computing for real parameter spaces. In *Stat Comput* 16, pp. 239–249. DOI: 10.1007/s11222-006-8769-1.

Bye (2013): Værradarer i Norge. Available online at http://stigbye.motocross.io/misc/ver/meteorologikart/varradarer/index.html, checked on 12/9/2021.

Büttner, György; Kosztra, Barbara; Soukup, Tomas; Sousa, Ana; Langanke, Tobias (2017): CLC2018 Technical Guidelines. EEA. Service Contract No 3436. Available online at https://land.copernicus.eu/user-corner/technical-library/clc2018technicalguidelines_final.pdf, checked on 12/13/2021.

Collier, Christopher; Hardaker, Paul (2004): Using Radar in Hydrometeorology. In *Peter Meischner (Ed.): Weather Radar: Principles and Advanced Applications. Berlin, Heidelberg: Springer Berlin Heidelberg*, pp. 115–129. DOI: 10.1007/978-3-662-05202-0_4.

Devak, Manjula; Dhanya, C. T. (2017): Sensitivity analysis of hydrological models: review and way forward. In *Journal of Water and Climate Change* 8 (4), pp. 557–575. DOI: 10.2166/wcc.2017.149.

EEA (1993): CORINE land cover. Guide technique. Luxembourg: Office des publ. officiellles des Communautés Européennes. ISBN: 92-826-2579-6 (Environnement, sécurité nucléaire et protection civile, 12585).

Einfalt, Thomas; Arnbjerg-Nielsen, Karsten; Golz, Claudia; Jensen, Niels-Einar; Quirmbach, Markus; Vaes, Guido; Vieux, Baxter (2004): Towards a roadmap for use of radar rainfall data in urban drainage. In *Journal of Hydrology* 299 (3-4), pp. 186–202. DOI: 10.1016/j.jhydrol.2004.08.004.

Elo, C. A. (2012): Correcting and quantifying radar data. Norwegian Meteorological Institute, Norway (Technical Report, 2). Available online at https://www.met.no/publikasjoner/met-report/met-report-2012, checked on 2/2/2022.

Essou, Gilles R.C.; Arsenault, Richard; Brissette, François P. (2016): Comparison of climate datasets for lumped hydrological modelling over the continental United States. In *Journal of Hydrology* 537, pp. 334–345. DOI: 10.1016/j.jhydrol.2016.03.063.

Fitsum, M. Woldemeskel; Bellie, Sivakumar; Ashish, Sharma (2013): Merging gauge and satellite rainfall with specification of associated uncertainty across Australia. In *Journal of Hydrology* 499, pp. 167–176. DOI: 10.1016/j.jhydrol.2013.06.039.

Foehn, Alain; García Hernández, Javier; Schaefli, Bettina; Cesare, Giovanni de (2018): Spatial interpolation of precipitation from multiple rain gauge networks and weather radar data for operational applications in Alpine catchments. In *Journal of Hydrology* 563, pp. 1092–1110. DOI: 10.1016/j.jhydrol.2018.05.027.

Frogner, Inger-Lise; Singleton, Andrew T.; Køltzow, Morten Ø.; Andrae, Ulf (2019): Convection-permitting ensembles: Challenges related to their design and use. In *Q.J.R. Meteorol. Soc.* 145 (S1), pp. 90–106. DOI: 10.1002/qj.3525.

Fujiyoshi, Y.; Endoh, T.; Yamada, T.; Tsuboki, K.; Tachibana, Y.; Wakahama, G. (1990): Determination of a Z-R Relationship for Snowfall Using a Radar and High Sensitivity Snow Gauges. In *American Meteorological Society* (Volume 29), pp. 147–152. DOI: 10.1175/1520-0450(1990)029<0147:DOARFS>2.0.CO;2.

Førland, E.; Allerup, P.; Dahlström, B.; Elomaa, E.; Jónsson, T. (1996): Manual for Operational Correction of Nordic Precipitation Data (DNMI-Reports, 24). Available online at https://www.met.no/publikasjoner/met-report/met-report-1996, checked on 2/3/2022.

Gekat, Frank; Meischner, Peter; Friedrich, Katja; Hagen, Martin; Koistinen, Jarmo; Michelson, Daniel B.; Huuskonen, Asko: The State of Weather Radar Operations, Networks and Products. In *Peter Meischner (Ed.): Weather Radar: Principles and Advanced Applications. Berlin, Heidelberg: Springer Berlin Heidelberg*, pp. 1–51. DOI: 10.1007/978-3-662-05202-0_1.

Germann, Urs; Joss, Jürg: Operational Measurement of Precipitation in Mountainous Terrain. In *Peter Meischner (Ed.): Weather Radar: Principles and Advanced Applications. Berlin, Heidelberg: Springer Berlin Heidelberg*, pp. 52–77. DOI: 10.1007/978-3-662-05202-0_2.

Gonchukov, Leonid; Bugaets, A.; Gartsman, Boris; Lee, Kwan (2018): Weather radar data for hydrological modelling: an application for south of Primorye region, Russia. DOI: 10.13140/RG.2.2.17008.30728.

Gottschalk, Lars; Jensen, Jørgen Lundager; Lundquist, Dan; Solantie, Reijo; Tollan, Arne (1979): Hydrologic Regions in the Nordic Countries. In *Nordic Hydrology* 10, pp. 273–286. DOI: 10.2166/nh.1979.0010.

Gupta, Hoshin V.; Kling, Harald; Yilmaz, Koray K.; Martinez, Guillermo F. (2009): Decomposition of the mean squared error and NSE performance criteria: Implications for improving hydrological modelling. In *Journal of Hydrology* 377, pp. 80–91. DOI: 10.1016/j.jhydrol.2009.08.003.

Hailegeorgis, Teklu T.; Alfredsen, Knut; Abdella, Yisak S.; Kolberg, Sjur (2016): Evaluation of storage–discharge relationships and recession analysis-based distributed hourly runoff simulation in large-scale, mountainous and snow-influenced catchment. In *Hydrological Sciences Journal* 61 (16), pp. 2872–2886. DOI: 10.1080/02626667.2016.1170939.

Hargreaves, George H.; ASCE, F.; Allen, Richard G. (2003): History and Evaluation of Hargreaves Evapotranspiration Equation. In *Journal of Irrigation and drainage engineering* 129 (1), pp. 53–63. DOI: 10.1061/(ASCE)0733-9437(2003)129:1(53).

Hargreaves, George H.; Samani, Zohrab A. (1985): Reference Crop Evapotranspiration from ambient air temperature. In *American Society of Agricultural Engineers*. DOI: 10.13031/2013.26773.

Hasan, M. M.; Sharma, A.; Johnson, F., Mariethoz, G.; Seed, A. (2016): Merging radar and in situ rainfall measurements: An assessment of different combination algorithms. In *Water Resources Research* 52 (10), pp. 8384–8398. DOI: 10.1002/2015WR018441.

Huang, Shaochun; Eisner, Stephanie; Magnusson, Jan Olof; Lussana, Cristian; Yang, Xue; Beldring, Stein (2019): Improvements of the spatially distributed hydrological modelling using the HBV model at 1 km resolution for Norway. In *Journal of Hydrology* 577, p. 123585. DOI: 10.1016/j.jhydrol.2019.03.051.

Hyndman, Rob J.; Koehler, Anne B. (2006): Another look at measures of forecast accuracy. In *International Journal of Forecasting* 22 (4), pp. 679–688. DOI: 10.1016/j.ijforecast.2006.03.001.

IMDi (2019): Integreringen i Naustdal (t.o.m. 2019) kommune. Available online at https://www.imdi.no/tall-og-statistikk/steder/K1433/befolkning/befolkning_hovedgruppe/historikk, updated on 1/1/2021, checked on 12/7/2021.

Joss, J.; Waldvogel, A.; Collier C.G. (1990): Precipitation Measurement and Hydrology. In *American Meteorological Society, Boston, MA*. DOI: 10.1007/978-1-935704-15-7_39.

Kalnay, Eugenia (2012): Atmospheric Modeling, Data Assimilation and Predictability. In *Cambridge University Press*. DOI: 10.1017/CBO9780511802270.

Kimura, Ryuji (2002): Numerical weather prediction. In *Journal of Wind Engineering and Industrial Aerodynamics* 90 (12), pp. 1403–1414. DOI: 10.1016/S0167-6105(02)00261-1.

Kirchner, James W. (2009): Catchments as simple dynamical systems: Catchment characterization, rainfall-runoff modeling, and doing hydrology backward. In *Water Resources Research* 45 (2). DOI: 10.1029/2008WR006912.

Knoben, W. J. M.; Freer, J. E.; Woods, R. A. (2019): Technical note: Inherent benchmark or not? Comparing Nash-Sutcliffe and Kling-Gupta efficiency scores. In *Hydrol. Earth Syst. Sci.* 23 (10), pp. 4323–4331. DOI: 10.5194/hess-23-4323-2019.

Krajewski, W. F.; Smith, J. A. (2002): Radar hydrology: rainfall estimation. In *Advances in Water Resources* 25 (8-12), pp. 1387–1394. DOI: 10.1016/S0309-1708(02)00062-3.

Krause, P.; Boyle, D. P.; Bäse, F. (2005): Comparison of different efficiency criteria for hydrological model assessment. In *Advances in Geosciences* 5, pp. 89–97. DOI: 10.5194/adgeo-5-89-2005.

Ledesma, J.; Futter, M. (2017): Gridded climate data products are an alternative to instrumental measurements as inputs to rainfall–runoff models. In *Hydrological Processes* 31 (18), pp. 3283–3293. DOI: 10.1002/hyp.11269.

Lengfeld, Katharina; Kirstetter, Pierre-Emmanuel; Fowler, Hayley J.; Yu, Jingjing; Becker, Andreas; Flamig, Zachary; Gourley, Jonathan (2020): Use of radar data for characterizing extreme precipitation at fine scales and short durations. In *Environmental Research Letters* 15 (8), p. 85003. DOI: 10.1088/1748-9326/ab98b4.

Lindström, Göran; Pers, Charlotta; Rosberg, Jörgen; Strömqvist, Johan; Arheimer, Berit (2010): Development and testing of the HYPE (Hydrological Predictions for the Environment) water quality model for different spatial scales. In *Hydrology Research* (41 (3-4)), pp. 295–319. DOI: 10.2166/nh.2010.007.

Liu, Zhaofei; Wang, Yamei; Xu, Zongxue; Duan, Qingyun (2017): Conceptual Hydrological Models. In *Duan Q., Pappenberger F., Thielen J., Wood A., Cloke H., Schaake J. (eds) Handbook of Hydrometeorological Ensemble Forecasting, Springer, Berlin, Heidelberg. ISBN: 978-3-642-40457-3*. DOI: 10.1007/978-3-642-40457-3_22-1.

Lussana, C.; Seierstad, I. A.; Nipen, T. N.; Cantarello, L. (2019a): Spatial interpolation of two-metre temperature over Norway based on the combination of numerical weather prediction ensembles and in situ observations. In *Quarterly Journal of the Royal Meteorological Society* 145, pp. 3626–3643. DOI: 10.1002/qj.3646.

Lussana, Cristian; Elo, Christoffer A.; Rønning, Snorre S. (2016a): KliNoGrid RR-Rad: combination of radar-derived precipitation fields and rain gauge observations (No. 19/2016).

Lussana, Cristian; Nipen, Thomas N.; Seierstad, Ivar A.; Elo, Christoffer A. (2021): Ensemble-based statistical interpolation with Gaussian anamorphosis for the spatial analysis of precipitation. In *Nonlin. Processes Geophys.* 28 (1), pp. 61–91. DOI: 10.5194/npg-28-61-2021.

Lussana, Cristian; Saloranta, Tuomo; Skaugen, Thomas; Magnusson, Jan; Tveito, Ole Einar; Andersen, Jess (2018): seNorge2 daily precipitation, an observational gridded dataset over Norway from 1957 to the present day. In *Earth System Science Data* 10, pp. 235–249. DOI: 10.5194/essd-10-235-2018.

Lussana, Cristian; Saloranta, Tuomo; Skaugen, Thomas; Magnussson, Jan; Tveito, Ole Einar; Andersen, Jess (2017): Evaluation of seNorge2, a conventional climatological datasets for snow- and hydrological modeling in Norway. In *ESSD*. DOI: 10.5194/essd-2017-64.

Lussana, Cristian; Tveito, Ole Einar; Dobler, Andreas; Tunheim, Ketil (2019b): seNorge_2018, daily precipitation, and temperature datasets over Norway. In *Earth System Science Data* 11, pp. 1531–1551. DOI: 10.5194/essd-11-1531-2019.

Lussana, Cristian; Tveito, Ole Einar; Uboldi, Francesco (2016b): seNorge v2.0, Temperature. In *METreport* No. 14, ISSN 2387-4201. Available online at https://www.researchgate.net/publication/310794415_seNorge2_An_observational_gridded_dataset_of_temperature_for_Norway, checked on 12/8/2021.

Lynch, Peter (2008): The origins of computer weather prediction and climate modeling. In *Journal of Computational Physics* 227 (7), pp. 3431–3444. DOI: 10.1016/j.jcp.2007.02.034.

Lynch, Peter (2014): The emergence of numerical weather prediction. Richardson's dream. First paperback edition. New York, NY, USA: Cambridge University Press. ISBN: 9781107414839.

Marshall, J. S.; Palmer, W. Mc K. (1948): The distribution of raindrops with size. In *Journal of Atmospheric Science* 5 (4), pp. 165–166. DOI: 10.1175/1520-0469(1948)005<0165:TDORWS>2.0.CO;2.

McCuen, Richard; Knight, Zachary; Cutter, A. (2006): Evaluation of the Nash–Sutcliffe Efficiency Index. In *Journal of Hydrologic Engineering - J HYDROL ENG* 11. DOI: 10.1061/(ASCE)1084-0699(2006)11:6(597).

Milad, Jajarmizadeh; Sobri, Harun; Mohsen, Salarpour (2012): A Review on Theoretical Consideration and Types of Models in Hydrology. In *Journal of Environmental Science and Technology* 5, pp. 249–261. DOI: 10.3923/jest.2012.249.261.

Monteith, J. L. (1965): Evaporation and environment. In *Symposia of the Society for Experimental Biology* 19, pp. 205–234. Available online at https://repository.rothamsted.ac.uk/item/8v5v7/evaporation-and-environment, checked on 3/15/2022.

Mu, Qiaozhen; Zhao, Maosheng; Running, Steven W. (2011): Improvements to a MODIS global terrestrial evapotranspiration algorithm. In *Remote Sensing of Environment* 115 (8), pp. 1781–1800. DOI: 10.1016/j.rse.2011.02.019.

Müller, H.; Haberlandt, U. (2018): Temporal rainfall disaggregation using a multiplicative cascade model for spatial application in urban hydrology. In *Journal of Hydrology* 556, pp. 847–864. DOI: 10.1016/j.jhydrol.2016.01.031.

Müller, M.; Homleid, M.; Ivarsson, K.; Køltzow, M. A. Ø.; Lindskog, M.; Midtbø, K. H. et al. (2017): AROME-MetCoOp: A Nordic Convective-Scale Operational Weather Prediction Model. In *American Meteorological Society* 32 (2), pp. 609–627. DOI: 10.1175/WAF-D-16-0099.1.

Nash, J. E.; Sutcliffe, J. V. (1970): River flow forecasting through conceptual models part I — A discussion of principles. In *Journal of Hydrology* 10 (3), pp. 282–290. DOI: 10.1016/0022-1694(70)90255-6.

Nipen, T. N.; Seierstad, I. A.; Lussana, C.; Kristiansen, J.; Hov, Ø. (2020): Adopting Citizen Observations in Operational Weather Prediction. In *Bulletin of the American Meteorological Society* 101.1, E43-E57. DOI: 10.1175/BAMS-D-18-0237.1.

Nipen, Thomas (2022): NWPdocs. Norwegian Meteorological Institute. Available online at https://github.com/metno/NWPdocs/wiki/, updated on 4/4/2022, checked on 4/10/2022.

Norwegian Ministry of Petroleum and Energy (2021): Electricity Production. Available online at https://energifaktanorge.no/en/norsk-energiforsyning/kraftproduksjon/, updated on 5/11/2021, checked on 2/4/2022.

NVE (2015a): Målinger og metoder. Available online at https://www.nve.no/vann-og-vassdrag/hydrologiske-data/vannstand-og-vannforing/malinger-og-metoder/, updated on 6/21/2021, checked on 3/10/2022.

NVE (2015b): Ofte stilte spørsmål. Available online at https://www.nve.no/vann-og-vassdrag/hydrologiske-data/vannstand-og-vannforing/malinger-og-metoder/, updated on 6/21/2021, checked on 3/10/2022.

NVE (2015c): Stasjonsnettet. Available online at https://www.nve.no/vann-og-vassdrag/hydrologiske-data/vannstand-og-vannforing/stasjonsnettet/, updated on 10/28/2021, checked on 3/10/2022.

NVE Atlas: Bre Area 1999-2006. Edited by Andreassen Peereboom. Available online at http://gis3.nve.no/link/?link=breatlas, checked on 12/7/2021.

Ojha, Chandra; Berndtsson, R.; Bhunya, P. K. (2008): Engineering hydrology: Oxford University Press. Available online at https://www.researchgate.net/publication/264895381_Engineering_hydrology, checked on 1/14/2022.

Orlowsky, Boris; Seneviratne, Sonia I. (2014): On the spatial representativeness of temporal dynamics at European weather stations. In *Int. J. Climatol.* 34 (10), pp. 3154–3160. DOI: 10.1002/joc.3903.

Pers, Charlotta (2022a): Forum discussion: Issue with extracting parameters from best simulations during autocalibration, 4/1/2022. Available online at https://sourceforge.net/p/hype/discussion/1818967/thread/2356376426/?limit=25#ae45, checked on 4/21/2022.

Pers, Charlotta (2022b): Forum discussion: hourly model parameters, 4/13/2022. Available online at https://sourceforge.net/p/hype/discussion/1818968/thread/48e6c98b32/?limit=25#f341, checked on 4/20/2022.

Peura, Markus; Koistinen, Jarmo; Hohti, Harri (2006): Quality information in processing weather radar data for varying user needs. In *Computer Science*. Available online at https://www.semanticscholar.org/paper/Quality-information-in-processing-weather-radar-for-Peura-Koistinen/5e856f47b551937109b208bc8052b521c1f51ba5, checked on 4/9/2022.

Poméon, Thomas; Wagner, Niklas; Furusho, Carina; Kollet, Stefan and Reinoso-Rondinel, Ricardo (2020): Performance of a PDE-Based Hydrologic Model in a Flash Flood Modeling Framework in Sparsely-Gauged Catchments. In *Water* 12, no. 8, p. 2157. DOI: 10.3390/w12082157.

Raghavan, S. (2003): Radar Meteorology: Springer Science+Business Media, B.V. (27).

Reinemo, Petter (2012): Simulering av urban flom ved bruk av data fra værradar. Master Thesis. NTNU, Norway. Institutt for Vann- og miljøteknikk. Available online at https://ntnuopen.ntnu.no/ntnu-xmlui/handle/11250/242223, checked on 2/4/2022.

Ryu, Soorok; Song, Joon Jin; Kim, Yongku; Jung, Sung-Hwa; Do, Younghae; Lee, GyuWon (2021): Spatial Interpolation of Gauge Measured Rainfall Using Compressed Sensing. In *Asia-Pacific J Atmos Sci* 57 (2), pp. 331–345. DOI: 10.1007/s13143-020-00200-7.

Saldaña Espinoza, Carolina Isabel (2022): Study of unregulated flow conditions in Norwegian rivers: Strategy for improving lake outflow using HYPE model. Master Thesis. University of Stuttgart, Stuttgart. Chair for Hydrology and Geohydrology.

Šálek, M.; Brezková, L.; Novák, P. (2006): The use of radar in hydrological modeling in the Czech Republic – case studies of flash floods. In *Nat. Hazards Earth Syst. Sci.* 6, pp. 229–236. DOI: 10.5194/nhess-6-229-2006.

Saltelli, Andrea; Annoni, Paola; Azzini, Ivano; Campolongo, Francesca; Ratto, Marco; Tarantola, Stefano (2010): Variance based sensitivity analysis of model output. Design and estimator for the total sensitivity index. In *Computer Physics Communications* 181 (2), pp. 259–270. DOI: 10.1016/j.cpc.2009.09.018.

Schönfelder, Lennart (2017): Performance assessment of the semidistributed hydrological model HYPE for central Norway. Master thesis, NTNU (Trondheim), University of Stuttgart. Available online at https://ntnuopen.ntnu.no/ntnu-xmlui/handle/11250/2454731, checked on 12/10/2021.

Schönfelder, Lennart (2019): Forum discussion: How to improve poor low-flow performance?, 11/26/2019. Available online at https://sourceforge.net/p/hype/discussion/1818967/thread/2356376426/?limit=25#ae45, checked on 4/22/2022.

Schönfelder, Lennart; Baclet (2022): A distributed hydrological model of Norway for environmental assessments. Manuscript in preparation.

Seity, Y.; Brousseau, P.; Malardel, S.; Hello, G.; Bénard, P.; Bouttier, F. et al. (2011): The AROME-France Convective-Scale Operational Model. In *Monthly Weather Review* 139 (3), pp. 976–991. DOI: 10.1175/2010MWR3425.1.

Sen, Zekai (2016): Spatial Modeling Principles in Earth Sciences. 2nd ed. Cham: Springer International Publishing. ISBN: 978-3-319-41756-1.

Sevruk, Boris (2004): Niederschlag als Wasserkreislauf. Theorie und Praxis der Niederschlagsmessung. Zürich: ISBN: 8096934376.

Shephard, Ronald W.; Färe, Rolf (1974): The Law of Diminishing Returns. In *Zeitschr. f. Nationalökonomie* 34, pp. 69–90. DOI: 10.1007/BF01289147.

Sitterson, Jan; Knightes, Chris; Parmar, Rajbir; Wolfe, Kurt; Avant, Brian; Muche, Muluken (2018): An Overview of Rainfall-Runoff Model Types. In *International Congress on Environmental Modelling and Software* 41. Available online at https://scholarsarchive.byu.edu/iemssconference/2018/Stream-C/41, checked on 1/14/2022.

Sivasubramaniam, K.; Alfredsen, K.; Rinde, T.; Sæther, B. (2020): Can model-based data products replace gauge data as input to the hydrological model? In *Hydrology Research* 51.2, pp. 188–201. DOI: 10.2166/nh.2020.076.

Sivasubramaniam, Kuganesan; Sharma, Ashish; Alfredsen, Knut (2018): Estimating Radar Precipitation in Cold Climates: The role of Air Temperature within a Nonparametric Framework. In *Hydrology and Earth System Sciences DIscussions*. DOI: 10.5194/hess-2018-351.

Sivasubramaniam, Kuganesan; Sharma, Ashish; Alfredsen, Knut (2019): Merging radar and gauge information within a dynamical model combination framework for precipitation estimation in cold climates. In *Environmental Modelling & Software* 119, pp. 99–110. DOI: 10.1016/j.envsoft.2019.05.013.

SMHI (2021): HYPE Model Documentation. Available online at http://www.smhi.net/hype/wiki/doku.php?id=start, checked on 11/9/2021.

Sorooshian, Soroosh (2008): Hydrological modelling and the water cycle. Coupling the atmospheric and hydrological models / edited by Soroosh Sorooshian … [et al.]. Berlin, London: Springer. ISBN: 9783540778424.

Statistics Norway (2021): 07459: Befolkning, etter region, statistikkvariabel og år. Statistikkbanken (ssb.no). Available online at https://www.ssb.no/statbank/table/07459/tableViewLayout1/, checked on 11/9/2021.

Storn, Rainer; Price, Kenneth (1995): Differential Evolution: A Simple and Efficient Adaptive Scheme for Global Optimization Over Continuous Spaces. In *Journal of Global Optimization* 23.

Syed, Kamran H.; Goodrich, David C.; Myers, Donald E.; Sorooshian, Soroosh (2003): Spatial characteristics of thunderstorm rainfall fields and their relation to runoff. In *Journal of Hydrology* 271 (1-4), pp. 1–21. DOI: 10.1016/S0022-1694(02)00311-6.

te Linde, A. H.; Aerts, J. C. J. H.; Hurkmans, R. T. W. L.; Eberle, M. (2008): Comparing model performance of two rainfall-runoff models in the Rhine basin using different atmospheric forcing data sets. In *Hydrology and Earth System Sciences* 12 (3), pp. 943–957. DOI: 10.5194/hess-12-943-2008,

Thornthwaite, C. W.; Mather, John R. (1951): The role of evapotranspiration in climate. In *Archiv für Meteorologie, Geophysik und Bioklimatologie, Serie B* 3, pp. 16–39. DOI: 10.1007/BF02242588.

Thorstad, E. B.; Økland, F.; Johnsen, B. O.; Naesje, T. F. (2003): Return migration of adult Atlantic salmon, Salmo salar, in relation to water diverted through a power station. In *Fisheries Management and Ecology* 10 (1), pp. 13–22. DOI: 10.1046/j.1365-2400.2003.00324.x.

Uboldi, Francesco; Lussana, Cristian; Salvati, Marta (2008): Three-dimensional spatial interpolation of surface meteorological observations from high-resolution local networks. In *Met. Apps* 15 (3), pp. 331–345. DOI: 10.1002/met.76.

van Ravenzwaaij, Don; Cassey, Pete; Brown, Scott D. (2018): A simple introduction to Markov Chain Monte–Carlo sampling. In *Psychonomic Bulletin & Review volume* 25, pp. 143–154. DOI: 10.3758/s13423-016-1015-8.

Vaze, J.; Post, D. A.; Chiew, F. H. S.; Perraud, J.; Teng, J.; Viney, N. R. (2011): Conceptual Rainfall–Runoff Model Performance with Different Spatial Rainfall Inputs. In *Journal of Hydrometeorology* 12 (5), pp. 1100–1112. DOI: 10.1175/2011JHM1340.1.

Wergen, W. (2002): Datenassimilation – ein Überblick. In *Promet (27, Heft 3/4) - Die neue Modelkette des DWD I*, pp. 142–149.

Wernli, H. (2011): Wetter, Chaos und probabilistische Wettervorhersagen. In *Promet (37, Heft 3/4) - Probabilistische Wettervorhersage*, pp. 3–11.

WMO (2008): WMO Guide to Meteorological Instruments and Methods of Observation. 7th ed. ISBN: 978-92-63-10008-5 (8).

Wolff, M. A.; Isaksen, K.; Petersen-Øverleir, A.; Ødemark, K.; Reitan, T.; Brækkan, R. (2015): Derivation of a new continuous adjustment function for correcting wind-induced loss of solid precipitation: results of a Norwegian field study. In *Hydrol. Earth Syst. Sci.* 19 (2), pp. 951–967. DOI: 10.5194/hess-19-951-2015.

Wood, S. J.; Jones, D. A.; Moore, R. J. (2000): Accuracy of rainfall measurement for scales of hydrological interest. In *Hydrology and Earth System Sciences* 4, pp. 531–543. DOI: 10.5194/hess-4-531-2000.

Yang, Long; Smith, James A.; Wright, Daniel B.; Baeck, Mary Lynn; Villarini, G.; Tian, F.; Hu, H. (2013): Urbanization and Climate Change: An Examination of Nonstationarities in Urban Flooding. In *Journal of Hydrometeorology* 14 (16), pp. 1791–1809. DOI: 10.1175/JHM-D-12-095.1.

Yu, Wansik; Nakakita, Eiichi; Kim, Sunmin; Yamaguchi, Kosei (2015): Improvement of rainfall and flood forecasts by blending ensemble NWP rainfall with radar prediction considering orographic rainfall. In *Journal of Hydrology* 531, pp. 494–507. DOI: 10.1016/j.jhydrol.2015.04.055.

Zarei, A. R.; Zare, S.; Parsamehr, A. H. (2015): Comparison of Several Methods to Estimate Reference Evapotranspiration. In *West African Journal of Applied Ecology* 23 (2), pp. 17–25. Available online at https://www.ajol.info/index.php/wajae/article/view/130514, checked on 12/10/2021.

Zawadzki, J.; Cieszewski, C. J.; Zasada, Michał; Lowe, Roger; Zasada, C.; Lowe, M. et al. (2005): Applying geostatistics for investigations of forest ecosystems using remote sensing imagery. In *Silva Fennica*.