UPPSALA UNIVERSITET

SLU

UPTEC W 22015

Examensarbete 30 hp

Juni 2022

Evaluation of Streamflow Predictions in an Ungauged Swedish Catchment

A Study of Håga River

Hanna Pierrau

Civilingenjörsprogrammet i miljö- och vattenteknik

Master's Programme in Environmental and Water Engineering

Master's Thesis, 30 credits

Evaluation of Streamflow Predictions in an Ungauged Swedish Catchment

- A Study of Håga River -

Author
Hanna Pierrau

Supervisors
Reinert Huseby Karlsen
Jan Seibert

Subject Reader
Thomas Grabs

June 17, 2022

Abstract

Evaluation of Streamflow Predictions in Håga River - A Study of an Ungauged, Swedish, Catchment

Hanna Pierrau

The Håga river, located west of the Swedish city Uppsala, is currently without a proper gauging station. Knowing the streamflow is nonetheless important to, for example, be able to calculate the nutrient transport in the river. This project aimed to evaluate different indirect methods of streamflow estimation to investigate how they perform, in particular in relation to SMHI's S-HYPE model. Two of the methods used were based on transferring streamflow of nearby catchments to Håga, either by using relationships between the mean and standard deviation of the streamflow time series (MOVE), or by simply scaling relative to catchment size (DAR). Furthermore, a hydrological model, HBV, was calibrated for Håga using different amounts and types of calibration data. All the methods were then compared to streamflow data from a previously active gauging station in Håga.

It was found that the overall best method to estimate the streamflow in Håga was using the MOVE method with one particular donor catchment. However, the performance of the simpler MOVE and DAR methods varied a lot from catchment to catchment. HBV was found to be able to produce better performing simulations than S-HYPE, despite being a simpler model. Even HBV-calibrations using alternative or limited data could perform rather well, although rarely at the level of a calibration utilising all available streamflow data.

A big uncertainty of the study was the fact that the most recent available validation data for the Håga catchment was from two decades ago, when the old gauging station was decommissioned. Most likely the methods that worked well during the 90s would work well today as well, but this is a matter that could be studied further.

Keywords: Hydrology, Predictions in ungauged basins, Hydrological signatures, rainfall-runoff modelling, HBV, S-HYPE

Department of Earth Sciences, Program for Air, Water and Landscape Science, Uppsala University, Geocentrum, Villavägen 16, SE-752 36, Uppsala

Referat

Utvärdering av flödesuppskattningar i Hågaån - En studie av ett svenskt avrinningsområde utan mätsation

Hanna Pierrau

Hågaån, ett vattendrag som ligger väster om Uppsala, saknar i nuläget en mätstation för vattenföring. Att känna till flödet är dock ändå intressant, bland annat för att kunna beräkna näringstransporten i ån. Syftet med detta projekt var därmed att utvärdera och jämföra olika metoder för att uppskatta vattenflödet i Hågaån, särskilt för att undersöka hur de presterade i jämförelse med SMHI:s S-HYPE-modell. Två av metoderna som användes för detta baserades på att överföra flöden från närliggande vattendrag till Håga, antingen genom att använda förhållanden mellan medelvärde och standardavvikelse för flödesdatan (MOVE), eller genom att bara utgå från skillnader i områdenas storlek (DAR). Utöver det kalibrerades även den hydrologiska modellen HBV för Håga med olika typer och mängder av kalibreringsdata. Alla metoderna jämfördes sedan med data från en mätstation som tidigare funnits i Hågaån.

Resultaten visade att den över lag bästa metoden för att uppskatta flödet i Håga var MOVE-metoden i kombination med ett av de närliggande vattendragen. Hur väl dessa simplare MOVE- och DAR-metoder presterade varierade dock mycket beroende på vilket vattendrag som användes som donator. Det visade sig även att det gick att erhålla bättre resultat med HBV än de som gavs av S-HYPE, trots att HBV är en enklare modell. Även HBV-kalibreringar baserade på alternativ eller begränsad data kunde producera välpresterande simulationer, dock sällan på samma nivå som den kalibrering som använt all tillgänglig flödesdata.

En stor osäkerhet i projektet kretsar kring att den nyaste tillgängliga valideringsdatan från Hågaån var över två decennier gammal, då den mätstation som funnits stängdes ner. Med stor sannolikhet kommer metoderna som fungerade väl under 90-talet även fungera bra i modern tid, men detta är något som kräver vidare studier.

Nyckelord: hydrologi, uppskattningar i områden utan mätstation, hydrologiska signaturer, modellering, HBV, S-HYPE

Institutionen för geovetenskaper, Luft-, vatten- och landskapslära, Uppsala Universitet, Geocentrum, Villavägen 16, SE-752 36, Uppsala

Populärvetenskaplig sammanfattning

Trots att kännedom om flödet i ett vattendrag kan vara till stor användning är det sällan det faktiskt mäts. I en studie om Hågaån, ett vattendrag i Uppsala som numera saknar en mätstation för vattebflöde, visade det sig att den bästa metoden att uppskatta flödet (av de som undersöktes) var att med hjälp av en statistisk metod omvandla flödet från en annan å.

Att känna till flödet i ett vattendrag är viktigt av många olika anledningar; det kan användas till allt från beräkningar av hur mycket näring eller föroreningar som transporteras, till att vara ett underlag för beslut rörande vattenhantering. Tyvärr saknas ändå mätstationer i den absoluta majoriteten av världens alla vattendrag, och då måste andra strategier tas till för att få förståelse för hur vattenflödet i ett område beter sig.

I dagsläget uppskattas flödet i Sveriges vattendrag av en modell som heter S-HYPE. Modellen använder bland annat regn- och temperaturdata, tillsammans med information om det studerade området, för att uppskatta vattenflöden. Det finns dock väldigt många olika modeller och metoder för att uppskatta flöde, och en studie om Hågaån har undersökt hur flödet kan uppskattas bäst där.

En metod för flödesuppskattning är att omvandla mängden vatten som rinner i andra vattendrag på ett sätt så att de anpassas till Hågaån. Hur bra denna typ av metod fungerar visade sig bero mycket på vilken å som användes för att överföra flödet från. Hågaån hade dock tur nog att ha ett närliggande vattendrag som fungerade väldigt bra för detta syfte; till och med så pass bra att S-HYPE-modellen överträffades i uppskattningsförmåga.

Med "omvandlings-metoder" kan man dock bara uppskatta det historiska flödet, medan modeller ger möjligheten att undersöka flera olika omständigheter, exempelvis vad som skulle hända om det blev varmare eller kom mer regn. Att ha en bra hydrologisk modell kan alltså vara väldigt användbart. Därför utvärderades även hydrologiska modeller i studien. Det visade sig att en modell kallad HBV var bättre på att uppskatta flödet i Hågaån än S-HYPE, trots att HBV både har en enklare struktur och kräver mindre data. Trots detta finns inte alltid all den data som behövs för att kunna anpassa HBV-modellen till ett vattendrag ordentligt. Studien visade att HBV går att anpassas ganska bra även med begränsad data, men då fungerade modellen inte lika bra som S-HYPE.

Metoder och modeller åsido: om man verkligen vill ha bra koll på flödet bör det utnyttjas att det faktiskt sitter en anordning som mäter vattennivån i Hågaån. Det finns nämligen sätt att omvandla vattennivå till flöde. Trots att detta både är omständigt och har sina osäkerheter (vilket är varför de olika uppskattningsmettoderna ändå är användbara) kan nog inget, inte ens den bästa omvandlings-metod, antagligen mäta sig med faktisk data direkt från Hågaån - om en perkfekt tidsserie av flödet är vad man är ute efter.

Preface

This project concludes my five years at the Master's Programme in Environmental and Water Engineering at Uppsala University and the Swedish University of Agricultural Sciences. The Project was initiated by SLU and is a 30 credits Master's Thesis. Reinert Huseby Karlsen from SLU, Department of Aquatic Sciences and Assessment, has been the supervisor of the project, and Thomas Grabs from Uppsala University, Department of Earth Sciences, was the subject reader.

First and foremost I want to thank Jens Fölster at SLU for suggesting the project to begin with, making it possible for me to get to work with this subject. It has been very interesting, and I have really learn a lot! A big thank you is owed to both Reinert and Thomas as well, for giving valuable feedback and helping me move forward with the project.

Several other people have also been a great help and made it possible for me to finish this thesis. Most of all I want to thank Jan Seibert, University of Zurich, (who also acted as the assisting supervisor of the project) for taking the time to answer all my questions about the HBV model and how it works. I probably would not have been able to use the model properly without his aid. I am also very grateful to Mikel Östlund, SLU, who provided me with data from the Håga river as well as organised a trip to Håga so I got to try measuring the streamflow myself. Faruk Djodjic, SLU, who did the pilot study that inspired this project, and Claudia Teutschbein, Uppsala University, who provided me with the PTHBV-data, also deserve my thanks.

Lastly, I want to aim a thank you at all my friends at the W-program who have made these five years the best years of my life, constantly showering me with love and support; it would not have been the same without you! I also owe thanks to my family and other friends who have always been there for me when I needed them, especially during this project.

Hanna Pierrau
Uppsala, May 2022

Definitions

Catchment: The area of land which is drained to generate the water that flows through a certain point in the landscape. Synonymous with basin.

Donor catchment: A catchment from which data is taken and used to estimate, for example, streamflow in another catchment.

Ensemble mean: In this project the ensemble mean refers to combining several time-series of streamflow into one by taking the mean value of all series at each time step.

Hydrological signature: A hydrological signature is (often) a quantitative measure describing an isolated feature of hydrology, for example how quickly a stream responds to rainfall, or the average magnitude of high-flows in the stream.

Hydrological year: The hydrological year starts on October 1:st and ends on the last of September and is given the number of the year it ends in. The hydrological year 2022 thus starts in October 2021. Whenever years are mentioned in this project, it refers to the hydrological year.

Hydrograph: A graph that shows how the streamflow changes in relation to time.

Index methods: These are simple methods for estimating streamflow based on scaling some property of the flow from a donor catchment to fit the catchment of interest.

Streamflow: The volume of water that flows through a section of a river at a given time. Symbolised with the letter Q. Runoff and discharge are other words used to refer to streamflow, although the meanings can differ slightly.

Specific Discharge: The discharge generated per unit area of land. It can be calculated by dividing the streamflow of a river by the area of the river's catchment. For this project the unit mm/day has been used to measure specific discharge.

Rating curve: A curve showing the relationship between water-level and streamflow in a river. It can be used to convert measurements of water stage into discharge-values.

Regionalisation: When data is extrapolated from one, or several different, basins to make predictions in another catchment.

Contents

1 Introduction

Knowing how the flow of a river behaves is an important source of information when making water management decisions. However, streamflow is only monitored in a small fraction of rivers in the world, and most catchments remain completely ungauged (Blöschl et al. 2013). This is in part because the process of measuring streamflow requires resources for building and maintaining a gauging station, but even if such resources were available it is not always possible to measure the discharge in a satisfactory way.

This is true for areas like Håga, a nature reserve located in Uppsala, Sweden. The river running through the area, Håga river, lacks a good location for establishing a rating curve, and is currently not properly monitored. This is in part due to the area's status as a nature reserve which protects the land from being disturbed. Due to this protection trees and other sources of debris that can affect measurements cannot be removed, and instead branches and other scraps sometimes get stuck in the river, affecting the water-level. The relationship between discharge and water-level can thus vary over time, making it difficult to continuously measure the streamflow using a rating curve (Fölster 2022).

Despite the practical difficulties with gauging the Håga river, the streamflow is still of interest and today you can find modelled discharge data supplied by the Swedish Meteorological and Hydrological Institute's (SMHI) model S-HYPE (SMHI 2022a). The question is if this model is the best way to predict the streamflow of Håga river, or if other methods can produce better results - something that can be investigated by using historical data from an old SMHI gauging station. Investigating this is not only important as it would increase the accuracy of, for example, calculations on the transport of nutrients in the Håga river itself; it will also provide information on how streamflow predictions could be improved for other ungauged streams.

1.1 Purpose and research questions

The purpose of this project is to examine different ways of predicting the streamflow of Håga river. The following questions will be explored:

- How well do index methods, that adjust the discharge of nearby donor-streams to fit the streamflow of Håga, perform when estimating the runoff in Håga river?
  - Do different donor catchments and index methods have differing strengths regarding different hydrological conditions and signatures, both compared to each other and to S-HYPE?
  - What nearby stream gives the best estimations, and why?
- Can a calibrated HBV-model be used to produce better estimations than the S-HYPE model?
  - How does simulations from the HBV-model compare to S-HYPE and the index methods regarding different hydrological conditions and signatures?
  - How does the performance of the HBV-model differ depending on what data is used for calibration?
- Which metrics are most useful for the evaluations?

By investigating these questions an improved method of predicting the streamflow of Håga river can hopefully be found. This could in turn be used by interested parties as a more reliable source of predicted discharge values to, for example, base water management decisions on. Hopefully the project will also lead to gained insight about the value of different types of hydrological data when it comes to runoff predictions in the Håga river and other streams.

2 Theory

2.1 Predictions in Ungauged Basins

For most populated areas in the world, having knowledge of the streamflow of nearby rivers is important for a multitude of reasons. Runoff data can inform decision-makers of risks related to floods and droughts, so appropriate measures like designing drainage infrastructure and protecting ecosystems can be implemented. In addition to affecting the transportation of sediment, nutrients and other factors affecting water quality, streamflow also works as an indicator of water supply (Blöschl et al. 2013).

Runoff information is thus of great societal importance. The fact that the majority of catchments worldwide are ungauged (Blöschl et al. 2013) means that many parts of the world lack this information - assuming that direct measurements are the only way to get it. Fortunately, there exist methods for estimating the discharge in basins without direct runoff measurements, although no method, no matter how suitable or creative, is equal to using actual observed discharge data (ibid.).

Despite hydrological estimations having an importance to society, there are still no universal theories on how to estimate runoff at the catchment scale (Hrachowitz et al. 2013). This has, in part, to do with the complexity of the catchment as a system; in addition to involving many different mediums like rock, soil and vegetation, the hydrology of an area also varies with topography, climate and human manipulation. Hydrological processes also take place on many different scales in both time and space, and can involve different numbers, and combinations, of processes depending on the area (Blöschl et al. 2013).

Organising all the information from the vast number of studies examining different processes at different temporal and spatial scales, in different places in the world, affected by different environmental variables is thus, understandably, difficult. For this reason, no single method for runoff prediction exists, but instead literally hundreds of methods - all with their different strengths and weaknesses (Blöschl et al. 2013).

2.2 Methods for predicting runoff

There are several different aspects of streamflow that one might wish to predict, depending on the catchment and its surroundings. In some catchments knowing the annual runoff or the duration of low-flows might be what is most important, while the flood risk is the main interest in another catchment. There are multiple methods for predicting each of these hydrological signatures in isolation, which is one of the reasons so many methods for predicting runoff exists (Blöschl et al. 2013).

However, if the goal is to get a complete picture of the catchment and its dynamics it is necessary to try to estimate the entire hydrograph of the system; a continuous time-series of the river's discharge. The hydrograph includes information about low-flows and floods, as well as the timing of runoff peaks and much else. As it is the runoff signature that depicts the system behaviour in most detail, it is also the type of prediction that is most complex (Blöschl et al. 2013).

To be able to predict the streamflow of a catchment some sort of data is required to use as a basis for the prediction. In the case of completely ungauged basins that information cannot come from the catchment itself. Instead the data needs to be borrowed from another, gauged, catchment (a "donor catchment"), and then somehow be used to make predictions in the catchment of interest (Blöschl et al. 2013). This type of data extrapolation is called regionalisation and can according to He et al. (2011) be divided into two types: either an effort to directly transfer streamflow, or some other flow metric, from one catchment to another, or to use model parameters calibrated for the donor catchments in the catchment of interest.

For regionalisation efforts to be relevant there needs to be something that connects the two catchments. The simplest form of connection is spatial proximity, which is based on Toblers' first law of Geography: "*everything is related to everything else, but near things are more related than distant things*" (Tobler 1970). Although it, in the context of hydrological modelling, can be assumed that features that affect hydrology vary smoothly in space (Parajka et al. 2013), actual hydrological processes can vary a lot over short distances (Blöschl et al. 2013). However, it has been shown that the relation of runoff between nearby catchments is stronger in areas with a wetter climate (Patil et al. 2011).

An alternative way to choose a donor catchment is by catchment characteristics. This method assumes that catchments that have similar features also have similar hydrology. As there is a multitude of traits in a catchment that affect the hydrological processes it is common to use several characteristics when assessing the similarity of basins. Some catchment traits that often are used are land use, soil type and depth, topography, climate, and catchment area (Blöschl et al. 2013).

2.2.1 Index methods

Blöschl et al. (2013) define index methods as a type of statistical method for predicting streamflow, consisting of simple relationships between catchment characteristics, climate, and runoff, instead of physical laws like mass and energy balances. Index methods are a sub-category of these statistical methods, and are based on scaling some property of a catchment to fit the catchment of interest, in the case of this study the entire hydrograph (ibid.). The simplest way to do this is by re-scaling the runoff values by using the difference in area between the catchments (Hirsch 1979), but other methods exist as well, for example those that use the relation between the mean value and standard deviation of the flow of two catchments.

2.2.2 Hydrological modelling

The hydrological models used for hydrograph prediction are so-called rainfall-runoff models that use precipitation and other meteorological data to predict the discharge of a stream. The structure of these models can vary tremendously from simple water balance equations to complex physical models based on laboratory experiments, and everything in between. In comparison with index methods, hydrological models can be used for more than just recreating streamflow records. One such example is investigating how flow conditions would change if precipitation increased (Blöschl et al. 2013).

One category of hydrological models are the *conceptual models*, which aims to represent the main flow processes at the catchment scale. Usually, they consist of a limited number of water storage elements ("buckets") that are connected by fluxes (Blöschl et al. 2013). Even though the term "conceptual model" often is used in opposition to "physical model", this does not mean the conceptual models are disconnected from physical reality: they still follow physical laws such as the mass balance. The term "conceptual" can thus be somewhat misleading (Seibert et al. 2021).

These simple bucket-type models have been shown to work well for wet catchments, while dryer catchments, where a larger proportion of the rain evaporates, require more complex model structures (Atkinson et al. 2002). However, even for the wet systems, a big challenge remains if the system is ungauged: what data should be used to calibrate the model? This is an argument for using the more complex, physical models, as the parameters used to tune them, in theory, can be measured in the field. This is not the case for conceptual models that have parameters linked to how the catchment functions as a whole (Blöschl et al. 2013).

Calibration data is thus needed to find satisfactory parameter sets for a conceptual model. For gauged catchments using time-series of streamflow is the obvious choice, which of course is impossible in ungauged catchments. Instead, a hydrological model can be calibrated to one or more gauged donor catchments. The parameters can then in some way be transferred to the catchment of interest (Blöschl et al. 2013). In this type of situation it has been found that transferring whole sets of parameters works better than trying to combine or interpolate between parameters from several catchments (Bárdossy 2007). This is because there rarely exists a clear relationship between unique parameters and catchment characteristics, and the fact that parameters often are interdependent and can compensate for each other (ibid.).

Another possibility is to use limited or alternative hydrological data for calibration: even though long time-series of runoff does not exist there might be a few measurements of streamflow or data relating to the water-level of the river. One way of examining how well these calibration based on alternative data work for making predictions in an ungauged catchment is by analysing a catchment that is gauged, but treating it as if it was not by not using all the available data (Seibert et al. 2009; Seibert et al. 2016; Pool et al. 2019).

2.2.3 Studies using hydrological models

This method of using limited or alternative data from a gauged catchment is common and has been explored by several different studies. One example is Seibert et al. (2009) who used it to examine the value of only a few discharge measurements when calibrating a hydrological model. Although earlier studies (Perrin et al. 2007) had found that about 350 discharge measurements were needed to obtain robust hydrological models, Seibert et al. (2009) found that a quite small number of discharge observations taken within one year are enough to constrain a model. They found that model performance increased most between 2 to 16 observations, and that no additional increase in efficiency was seen after about 32 observations. Further findings were that an ensemble mean of well-performing models (combined through a weighted mean), as a rule, performed better than the single best-performing simulations (ibid.).

In another study a similar method, i.e. treating gauged catchments as ungauged, was used to study how well a hydrological model performs if stream water-level observations are used to calibrate it, instead of streamflow (Seibert et al. 2016). As water-level measurements only give information about the dynamics of the system, the volume of discharge was not evaluated in this type of calibration. The results showed that even though the water-level calibrations always had lower efficiency than models calibrated with complete streamflow data, water-level observations alone could lead to good model fits and constrain the model parameters. The method was found to work best for wet catchments, where runoff volumes largely are controlled by the rainfall input, in contrast to arid basins where more data was needed (Seibert et al. 2016).

It has also been shown that combining different types of hydrological data for calibration can be a way to improve the performance of hydrological simulations (Pool et al. 2021). Using both continuous water-level measurements as well as discharge point measurements while calibrating resulted in better models than using either type of data on its own. This was because the different data types supplied different information. While the continuous stream level data revealed the dynamics of the system, discharge observations helped to link the dynamics to the river's annual water balance, resulting in more correct discharge volumes (ibid.).

Pool et al. (ibid.) also found that when using discharge observations for calibration the timing of the measurements matter. Although as few as two to six measurements typically would increase the performance of the model, using only a small number of observations could sometimes result in worse simulations. This was due to the fact that some observations were more informative than others. The timing of the samples thus proved important, where data from discharge peaks and recession periods were the most useful (ibid.). It has also been found that the year one takes the measurements matters as well. During an informative year no more than three streamflow observations were needed, while the performance of the model increased with an increasing number of point measurements for less informative years (Pool et al. 2019).

2.3 Measures of Performance

Performance measures are used to quantify how well a method estimates the discharge of a river by comparing the modelled streamflow with the actual observed streamflow. There are multiple ways this can be done.

2.3.1 Goodness-of-fit measures

One way is by using so-called goodness-of-fit functions that compare the estimated and observed streamflow by calculating a value based on the entire streamflow time-series. The similarity between the two time-series is then described with a value, where 1 often indicates a perfect fit. Apart from evaluating how well a method predicts runoff goodness-of-fit measures are also used in the calibration of hydrological models. In this case the objective is to minimise the error between simulation and observed values by getting the goodness-of-fit measure, also called objective function, as close to its optimal value as possible (Hrachowitz et al. 2013). Unfortunately, there is no one objective function that can give a simple, comparable, overall evaluation of how well the hydrograph is approximated. Instead, different objective functions prioritise different parts of the hydrograph, and subsequently what is deemed to be a "good" estimation can differ between two goodness-of-fit measures. The choice of objective function (see also Section 3.5.1) thus becomes very important as it affects the value of the parameters and hence also the model results (Sorooshian et al. 1983).

2.3.2 Hydrological signatures

In addition to goodness-of-fit measures, that summarise the comparison between the entire hydrographs into one single number, it has become increasingly popular to look at different aspects of hydrology separately (see also Section 3.5.2). These "aspects", called hydrological signatures, are represented by specific parts of the hydrograph and can, among other things, be used to evaluate model performance (McMillan 2021).

Hydrological signatures are values, often scalars, that describe individual features of hydrological behaviour, like what type of flows are common for a stream, or how the catchment responds to a change of weather (Figure 1). The signatures can be derived from modelled or observed hydrological data such as soil moisture or precipitation, although streamflow is most commonly used (McMillan 2020). The complexity of the signatures varies from simple statistics like mean flow values, to more complex metrics that describe the source of the water in the stream (McMillan 2021).

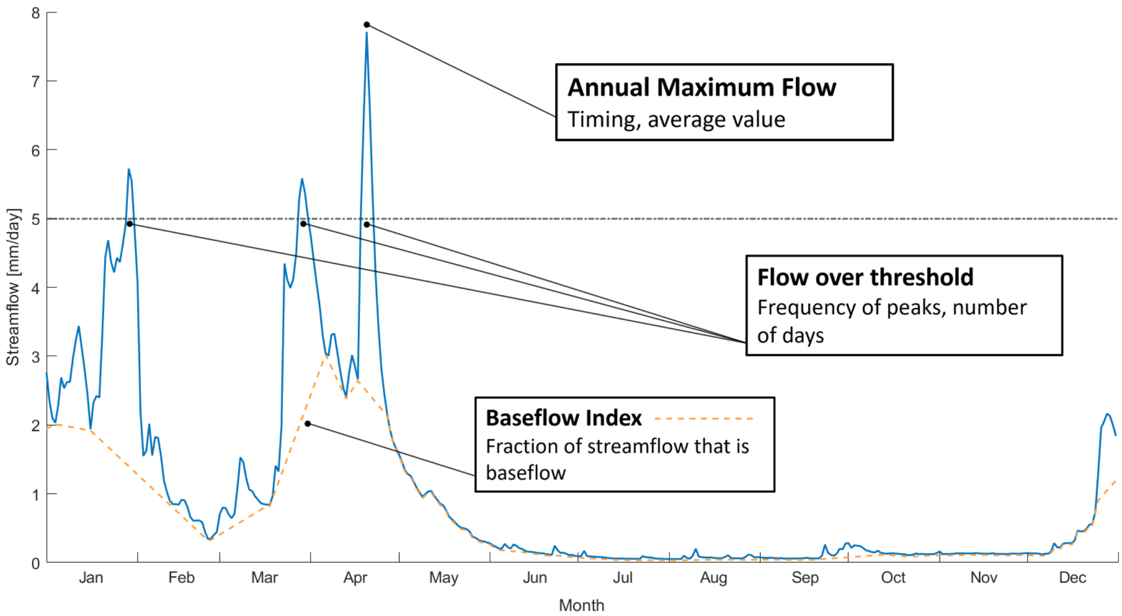

Figure 1. *Some examples of hydrological signatures that can be calculated from different types of data*

Hydrological signatures have been used for many different purposes across a wide variety of hydrological fields for a long time. Thus an abundance of hydrological signatures exists, and more are being created (Addor et al. 2018). Due to this profuse amount of signatures there are many overlaps, such as indices for the duration of low-flow events using different thresholds to define what "low flow" is (McMillan 2021). It is also very common for different signatures to be intercorrelated despite pertaining to different hydrological behaviours. A study by Olden et al. (2003) analysed 171 different hydrological signatures and found that up to four indices often were enough to describe the dominant patterns in the catchments due to the correlation between the different signatures.

3 Material and methods

3.1 Scope and delimitations

This project is limited to evaluating the performance of two different index methods and two hydrological models for estimating the streamflow of Håga river. As there only exists recorded streamflow for Håga river between 1979 and 2001 all streamflow estimations were done for the ten hydrological years 1991 to 2000, so the methods could be validated. The ten preceding years (1981-1990) were used for calibration.

The donor catchments evaluated in the project were limited to the five nearby catchments with rivers that were gauged during the same time the Håga river was gauged.

Although it is not a strict limitation this project focuses somewhat more on the performance of the different methods during high- and low-flows in comparison to other flow conditions. This is due to an interest in high-flows for more accurate predictions of the transport of nutrients like phosphor, and because low-flows are of ecological importance (Fölster 2022).

3.2 Description of the study area

The Håga river (Figure 2) is located to the west of the city Uppsala, in the east of central Sweden, and has a catchment area of about 122 km^2 (SMHI 2022a). The river drains the lake Fibysjön, located just northwest of Vänge, and then flows southeast through a landscape mainly dominated by forests and agricultural land (Figure 3), including the nature reserve Hågadalen-Nåsten (Persson et al. 2011). The Håga river is strongly affected by morphological changes connected to the use and drainage of the land around the river, and is, after the Fyris river, the second largest inflow to the lake Ekoln (Länsstyrelsen Västmanlands län n.d.).

Figure 2. *A photography of the Håga river showing how the it can get dammed up. The picture is taken a few hundred meters upstream of the old SMHI gauging station, and a few dozen meters downstream of where a water-level gauge currently is placed.*

The average yearly precipitation in the catchment is around 600 mm, approximately a fourth of it falling as snow. About half of it evaporates, letting around 300 mm of water per year turn into runoff; water that eventually ends up in the Håga river. Between 1979-2001 there existed an SMHI-station measuring discharge in the river (SMHI 2022b). The gauge was decommissioned due to economic reasons, as the station most likely needed a great deal of maintenance, which was considered too expensive (Sandehed 2022). In 2016 a gauge for continuous water-level measurements was installed, and there are plans to develop a rating curve for the Håga river (Östlund 2022).

In addition to Håga itself, five gauged catchments in the proximity of Håga were chosen to be evaluated as potential donor catchments (Figure 3). They range in size between 6-721 km^2 and are like Håga characterised by quite small elevation differences, and have forests as the dominating land cover.

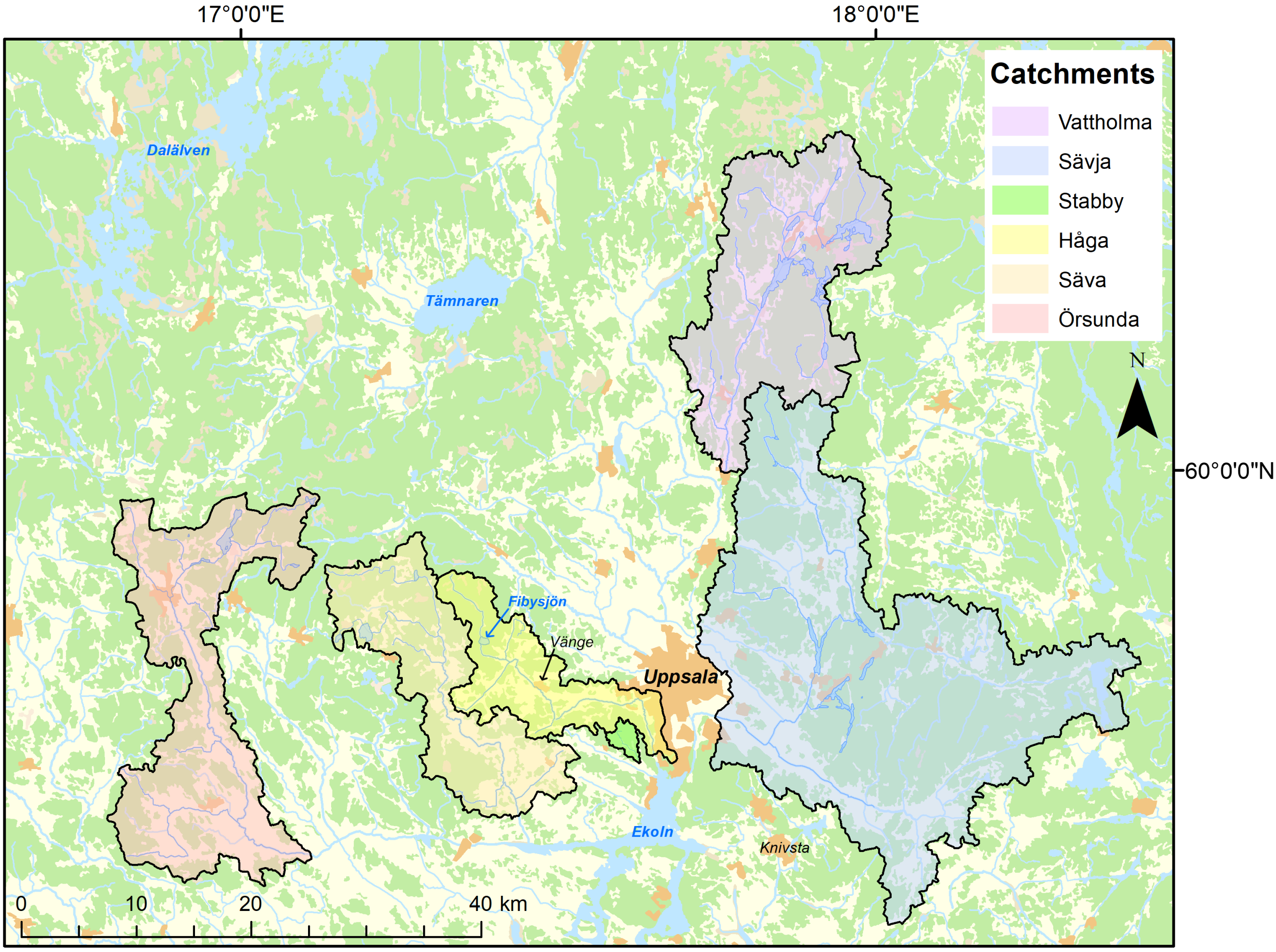

Figure 3. *The catchments chosen for the study. Background map ©Lantmäteriet*

Säva and Sävja are the two basins that are most similar to Håga when looking at the traits listed in, while Stabby is the catchment with the traits that are most dissimilar to Håga (Table 1). The data in Table 1 mainly comes from SMHI's service *Modelldata per område* (SMHI 2022a), but average soil depth and mean slope was calculated using GIS-analyses.

Table 1. Some catchment traits of Håga and the donor catchments. Paved surfaces include both urban areas and the category "hard surfaces" from SMHI's *Modelldata per område*. The "missing" areal percentages consist of open land. The darker green indicates the catchment that is most similar to Håga, while the lighter green shows the second best catchment. The red colour indicates the catchment most dissimilar to Håga in that category. Note that some values might be very similar despite having cells with different colours.

Catchment	Catchment area [km^2]	Average soil depth [m]	Mean slope [%]	Forest [%]	Lake [%]	Paved surfaces [%]
Håga	122	3.8	3.1	61	0.40	7.6
Säva	197	4.4	2.8	64	0.9	1.2
Stabby	6.18	2.4	2.6	87	0.03	0.1
Sävja	721	3.8	3.0	63	1.6	1.7
Vattholma	294	4.1	1.9	74	4.0	1.9
Örsunda	312	7.5	2.6	54	1.0	2.2

When in comes to precipitation the catchments are also very similar. However, the runoff ratio (the amount of precipitation that becomes runoff) in Håga is higher than in the other catchments (Appendix 8.1).

3.3 Data

Several different sets of data were used in this project. Streamflow data for the six catchments of interest was obtained from SMHI's gauging stations. Simulated streamflow data for Håga river from the S-HYPE model was also downloaded from SMHI (SMHI 2022a).

Another SMHI-supplied data set that was used was precipitation and temperature data from the PTHBV-database. This is a database especially aimed toward calculations with hydrological models, the HBV-model in particular. The data consists of values interpolated between SMHI's meteorologic measuring stations into a nationwide grid with a resolution of 4x4 km. The interpolation is aided by information on topography and wind, and the precipitation data is corrected for measurement errors (SMHI n.d.).

Lastly some data was gathered directly from the Håga river. This data consists of measurements of the water stage in the river collected by a TruTrack capacitance water-level logger (TruTrack, Inc., model WT-HR) installed by SLU in April 2016 (Östlund 2022). The water-level is measured continuously and recorded on a 30-minute basis. Due to the finite amount of storage in the logger there are gaps in the data between 2016-12-27 to 2017-01-19, 2018-12-19 to 2018-12-27, 2019-09-23 to 2019-10-25 and 2020-08-12 to 2020-08-24. This data was used to create daily mean values of the water-level, which were used for the analyses. Although the water-level gauge is placed in a location where the water stage normally only is affected by the upstream conditions, in February 2020 some trees fell into the river and hindered the water from flowing freely (Östlund 2022).

Furthermore, the streamflow was manually measured on three occasions in 2018 using a current meter. The three measurements cover a range from low flows (0.015 mm/day) to high flows (1.92 mm/day), and also includes a measurement of an intermediate flow (0.72 mm/day) (Appendix 8.2). One additional measurement was made during this project, but at a point in time that was too late to include it in the data analysis.

3.4 Methods for predicting runoff

3.4.1 Index methods

The simplest form of index method for estimating a discharge time-series is to assume that the only factor that makes the discharge of two rivers differ is the difference in catchment area. This is called the Drainage Area Ratio (DAR) method (Hirsch 1979) and is thus based on the assumption that the runoff per unit area, also called specific runoff, of the two catchments is the same (Equation 8).

$$Specific\ runoff = \frac{Streamflow}{Catchment\ area} \tag{1}$$

An alternative way to transfer runoff data from a donor catchment is to use the Maintenance of Variance (MOVE) method (Hirsch 1982). Instead of standardising by area this method assumes that it is the mean and variance that differ between the runoff series of the two catchments (Equation 2):

$$Q_I(i) = \mu_I + \frac{\sigma_I}{\sigma_D}(Q_D(i) - \mu_D), \tag{2}$$

where Q_I is the streamflow in the catchment of interest, Q_D is the flow in the donor catchment while μ and σ are their respective mean flow values and standard deviation of the flow (for the method to work better the variables can be calculated using a log-transformed streamflow time series). The runoff values of the ungauged catchment are in this way computed by the difference in mean and variance between the two areas. This method thus requires some runoff data from the catchment of interest, and it is, therefore, a so-called "record extension technique" (ibid.).

What these two methods have in common is that they use runoff from another catchment to decide both the magnitude and the timing of the streamflow in the ungauged catchment. Using catchments that are closely located thus become important, as the timing of rainfalls, and subsequent runoff peaks, otherwise would differ too much (Blöschl et al. 2013).

3.4.2 Hydrological Models

In addition to the index methods this project compared two different hydrological models, HBV and S-HYPE. While the HBV-model was calibrated in several different ways specifically for this project, S-HYPE was only used in the sense that the data already simulated by SMHI was downloaded and compared to the other methods.

HBV

The HBV-model is a conceptual model named after the research unit at SMHI that created it, Hydrologiska Byråns Vattenbalansavdelning. It can both be run in a lumped way, treating the entire catchment as one homogeneous unit (the case for this project), or in a semi-distributed way, by dividing the basin into different elevation and vegetation zones (Seibert et al. 2021). The model consists of four main routines (Figure 4) that work together to simulate the discharge of a modelled catchment from precipitation, temperature and potential evaporation data (Bergström 1990).

The first routine is a snow routine which regulates precipitation and whether it falls as snow or rain, as well as the accumulation, melting and refreezing of snow. The water that is generated by this routine is then passed on to the soil routine (Bergström 1990). The soil routine is highly non-linear and determines how much of the precipitation contributes to the generation of runoff in the next routine, and how much gets stored in the soil box or evaporated. The emptier the soil box is, the more water stays in it (ibid.).

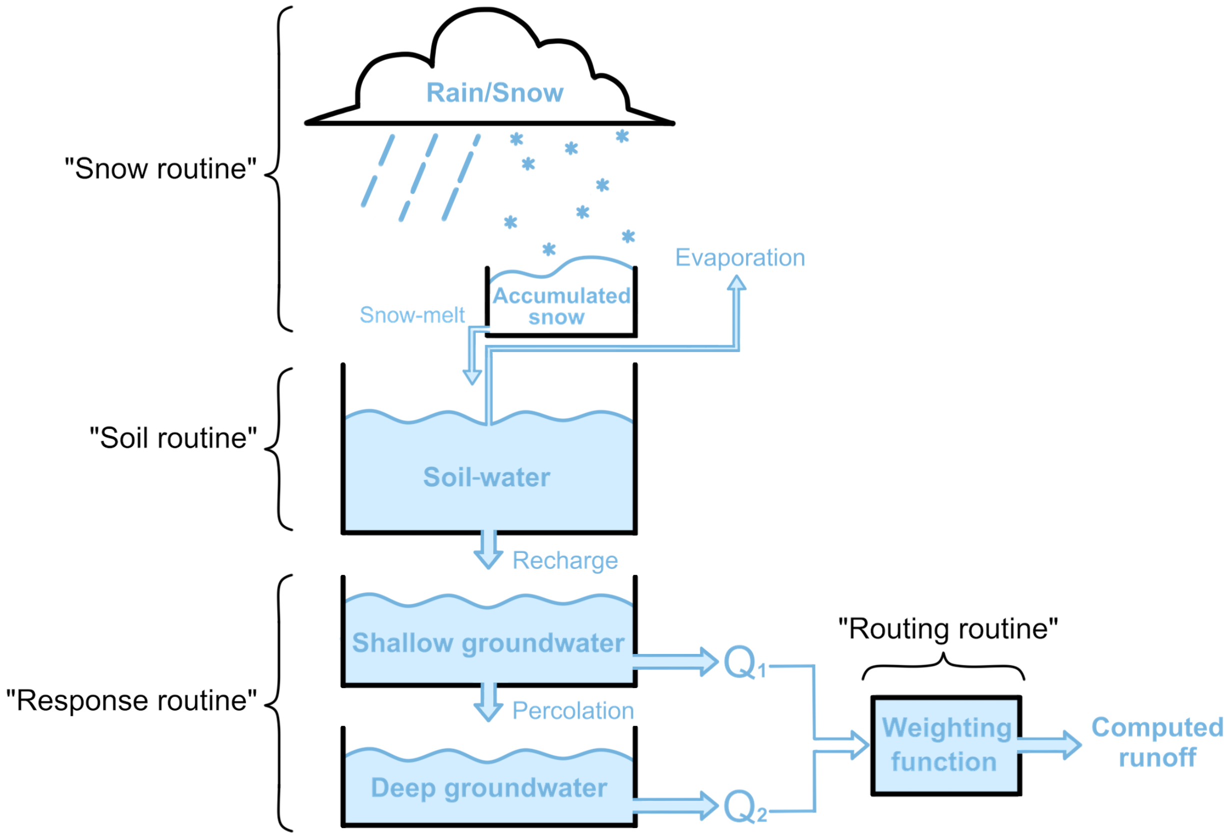

Figure 4. *A very simplified version of the HBV-model's structure.*

The third routine is the groundwater routine, also called the response function. There are several different versions of this routine, but the one used in this project consists of two boxes with one outlet from each: an upper reservoir corresponding to more shallow groundwater, involved in faster runoff response and peak flows, and a lower reservoir corresponding to the deep groundwater which feeds the baseflow of the river (Lindström et al. 1997). The combined outflow from the two groundwater-boxes is then passed to the last routine called the routing routine. This consists of a weighting function which is used to simulate the effect that water from different areas in a catchment reaches the outlet after different amounts of time (Häggström 1989).

Each of the routines in HBV is controlled by a set of parameters, and the values of these parameters are obtained through model calibration (Bergström 1990). A study by Seibert et al. (2016) has compiled suitable parameter values for the model (Appendix 8.3).

Today the HBV model has existed for around 50 years, and several different versions of the model exist (Seibert et al. 2021). One such version is HBV-light, a version that has been further developed to include functionalities like automatic calibration and different model variants (ibid.). HBV-light is the version of the model that was used in this project.

S-HYPE

The HYPE (Hydrological Predictions for the Environment) model is an open-source hydrological model that was developed by SMHI from 2005 to 2007. The reason behind its development was that SMHI wanted to create a tool with a high spatial resolution that could be used to calculate the transport of substances in a more effective way than the version of HBV that was used at the time. HYPE was built based on experiences from using the HBV-model, and the two models thus share some similarities (Lindström et al. 2010).

Both models are conceptual and semi-distributed, but differ in the way that the parameters of the HYPE-model are connected to soil type and land use. Thus the HYPE-model is also somewhat more physically based, as some of the equations of the model are connected to physical properties like the porosity and infiltration capacity of the soil (Gustavsson 2014). The HYPE model also requires more input data than HBV; in addition to meteorological data (which HBV also needs), it also needs information about soil types, land use and the depth of lakes in the area, among other things (Lindström et al. 2010).

S-HYPE is the version of HYPE built to encompass the entirety of Sweden. S-HYPE is continuously improved upon by SMHI when it comes to new input data, ways to describe processes as well as calibration of the model (SMHI 2020). The HYPE-model has a large number of parameters that need to be calibrated, which for S-HYPE is done in two steps. Firstly a general "base model" is produced by using data from many different gauging stations. The base model is thus a compromise between several areas, and not adapted to any single station. The second step is to make local adaptions of the model, which is only done for a couple of key parameters (SMHI 2018).

3.5 Calibration and evaluation metrics

3.5.1 Calibration

To fit different calibration situations, using different data-sets, three different objective functions were used when calibrating HBV. One of them is the Kling-Gupta efficiency (KGE) (Gupta et al. 2009):

$$KGE = 1 - \sqrt{(r-1)^2 + (\frac{\sigma_{sim}}{\sigma_{obs}})^2 + (\frac{\mu_{sim}}{\mu_{obs}})^2}, \tag{3}$$

where r is the correlation between the observed and simulated discharge time-series, σ is the standard deviation and μ the mean discharge of the two time-series. KGE has a tendency to prioritise high flows over low flows. Despite this it still often leads to underestimations of peak flows, although this error is smaller than for other, more traditional, objective functions(ibid.). Due to this KGE has been used increasingly since its development (Knoben et al. 2019).

As KGE uses mean values and standard deviations it is not as suitable as a calibration metric when there are very few streamflow observations available. In these cases an objective function pre-defined in HBV was used, which is based on the mean squared error (MSE) (Equation 4)

$$MSE = \frac{\sum_{t=1}^{n}(Q_{obs} - Q_{sim})^2}{n}, \tag{4}$$

but also normalises by the flow and is formulated in a way that makes a value of 1 mean a perfect fit (Vis 2022). The objective function, called *ReffQObsSample*, is formulated as seen in Equation 5.

$$ReffQObsSample = 1 - \frac{1}{m}\sum \frac{\frac{1}{n}\sum(Q_{obs} - Q_{sim})^2}{{Q_{obs}}^2} \tag{5}$$

Where m is the number och streamflow observations, and n the number of days with the same water level as when one of the observations were measured (thus water level data is also required to use this function).

KGE and *ReffQObsSample* both measure the performance of the model using discharge values. In cases where the water stage is used to calibrate the hydrological model such objective functions cannot be used as stream level holds no information about the volumes of water being transported. However, the rise and fall of streamflow and water-level in a river are correlated and the water stage can thus be used for calibration to capture the dynamics of the system. Seibert et al. (2016) found that using the Spearman rank correlation coefficient (Spearman 1904), a non-parametric correlation measure, as objective function could be used to accomplish this. Good fits, values close to one, are achieved when stream level and streamflow are monotonically related, but do not require that the actual values or the exact shape of the two curves are the same. Spearman rank was thus used as objective function when HBV was calibrated with water level data.

3.5.2 Evaluation

The evaluation of the performance of the different streamflow prediction methods was conducted mainly by comparing the hydrological signatures listed below. However, to get an overall assessment of performance KGE and Spearman rank were used as evaluation metrics as well, the latter to inform how well the timing of the runoff peaks was captured by the different methods.

Flow Duration Curve
A Flow duration curve (FDC) is a graph showing how often a certain magnitude of streamflow is equalled or exceeded. As the FDC condenses a large amount of information about streamflow variability into a single image, and due to the usefulness of such information for both managing water resources and environmental health, FDCs are used for a wide range of purposes (Blöschl et al. 2013). They are also commonly used in hydrological studies to get an evaluation of a catchment's "overall behaviour", in addition to characterising the duration and magnitude of certain flows (McMillan et al. 2017).

However, by representing the streamflow in the frequency domain information on timing is lost (Blöschl et al. 2013). The FDC also combines effects of different hydrological processes due to the fact that the same flow value, for example, can occur as both as peak value or a recession of a larger flow (McMillan et al. 2017).

Water Balance
To get a general indicator of how well the overall water balance of the system is captured

by the different methods the difference in the generated volumes of water (per year) between the estimations and reality was calculated:

$$Q_{diff} = \frac{\sum_{t=1}^{n} V_{est,\ t} - \sum_{t=1}^{n} V_{obs,\ t}}{\Delta years} \tag{6}$$

Where $V_{est,\ t}$ is the estimated volume of water that passes through a stream for day t , $V_{obs,\ t}$ is the observed daily volume in Håga river, both measured in [mm] in this study. The variable $\Delta years$ is the number of years of streamflow data used for the calculation, giving Q_{diff} the unit [mm/year].

Flashiness Index
The amount of change in discharge volumes in a river can be described by a flashiness index (FI), which expresses changes in discharge magnitudes relative to the total streamflow (Equation 7).

$$FI = \frac{\sum_{t=1}^{n} |Q_t - Q_{t-1}|}{\sum_{t=1}^{n} Q_t} \tag{7}$$

Flashiness can be described as the rate of change in the flow of a stream, and a high FI thus corresponds to a hydrograph with rapid fluctuations, while a low value indicates a stream that changes more slowly (Baker et al. 2004).

Baseflow index
The baseflow of a catchment is the part of the streamflow that comes from water that has been stored deep in the ground (Blöschl et al. 2013). The baseflow is thus strongly related to the amount of groundwater in the area, and a stable baseflow implies large groundwater storage and that more water takes long flow-paths through the catchment (McMillan 2020).

The baseflow index (BFI) is defined as the proportion of the total streamflow that is made up of baseflow, and it can be calculated in several different ways (McMillan 2020). One commonly used method, and the one used in this project, is the method described by Gustard et al. (1992), which in essence calculates the baseflow through an algorithm that smooths out the hydrograph (Figure 1).

Signatures of high- and low-flows
Many different types of hydrological signatures can be used to analyse high- and low-flows, in this project three types of signatures have been used. The first signature was the bias between the estimated and observed flows for different flow quantiles representing high- or low-flow (Equation 8).

$$Bias_q = \frac{\sum_{t=1}^{n} (Q_{t,q}^{est} - Q_{t,q}^{obs})}{\sum_{t=1}^{n} Q_{t,q}^{obs}} \tag{8}$$

In the equation q represents the quantile used. In this project Q90 and Q95 (the 90th and 95th quantiles, that is runoff values that are exceeded 90 or 95 % of the time) for Håga river were used to represent low-flows, and Q10 and Q5 represented high-flows. A bias equal to 0 means a perfect fit, which requires the method to both have the right magnitude and timing of the flow.

The second type of signature was the frequency of high- and low-flow days, calculated by finding the average number of days per year where the flow surpassed or was below a certain threshold. In this project the thresholds were set to five times the median daily flow for Håga river for high-flow, and the median daily flow divided by five for low-flow. These thresholds were chosen as they resulted in a reasonable number of days of high- and low-flow each year.

Lastly the mean high-flow (MHQ) and mean low-flow (MLQ) were calculated by taking the average of the annual maximum or minimum flow. These signatures were studied as they are standard signatures used by SMHI, in addition to giving a general indication of flow magnitudes.

3.6 Method

3.6.1 Index methods

The two Index methods DAR and MOVE were used to estimate the runoff in Håga river for the evaluation period (1991-2000), resulting in 10 separate estimations of the flow in Håga: one estimation for each index method for the five different donor catchments. For the DAR method this was simply done by calculating the specific discharge of the different catchments, i.e. dividing the discharge of the streams by the area of the catchment, so the flow of the streams could be compared.

To use the MOVE-method a few more calculations were required. The formula (Equation 2) used to estimate the discharge in the stream of interest requires standard deviation and the mean value of the flow in both Håga river and the river of the donor catchment. These values were calculated with data from the years 1981-1990, the ten years preceding the estimation period. As discharge, as a rule, is not normally distributed it was decided to use the logarithms of the streamflow data for all calculations, and take the anti-logarithm of the values after the calculations were done. This was done in accordance with how the method was used by Hirsch (1982), and it has been proved to produce better results.

3.6.2 Hydrological models

The HBV-model was calibrated in six different ways using different data. As with the index method all streamflow values were estimated (simulated) for the period 1991-2000. However, what years that were used to calibrate the model depended on the data used for the calibration, as the different data sets not always were available for the same years. In all cases the model was warmed up using the three hydrological years that preceded whatever years were used for the calibration or simulation.

In addition to precipitation and temperature data, the HBV-model requires potential evaporation to run. The evaporation was calculated using the formula (Equation 9) suggested by Oudin et al. (2005):

$$PE = \frac{R_e}{\lambda\rho}\frac{T_a + 5}{100} \tag{9}$$

where PE is the rate of potential evapotranspiration (mm/day), R_e is extraterrestrial radiation(MJ/m^2 per day), λ is the latent heat flux in (MJ/kg), ρ is the density of water (kg/m^3) and T_a is mean daily air temperature (°C), derived from long-term mean values.

If $T_a + 5$ is smaller than 0, PE is set to zero. This formula was used as Oudin et al. (2005) found that it led to models performing somewhat better than when using the traditional, more complicated, Penman-Monteith formula.

For all calibrations the overall procedure was the same, and all the parameter sets were optimised within the same predefined range for each parameter (Appendix 8.3). This was done by using a genetic algorithm that, through stochastic processes, selects and recombines parameter sets from previously existing sets (in a way that reflects natural evolution) to find optimal parameter values (Seibert 2000). For each data set used for calibration genetic algorithm was used to obtain 100 optimised parameter sets to account for parameter uncertainty. This was done as the genetic algorithm is based on stochastic elements and the same input thus results in different parameter sets each time. The resulting hydrographs from these 100 parameter sets were combined into one discharge time-series by taking the mean value of each time step, creating one ensemble mean that was analysed further.

The HBV-model was calibrated in six main ways, each way using different sets of calibration data. First and foremost, the model was calibrated with ten years (1981-1990) of daily discharge data from the Håga river, using KGE as the objective function. This calibration was used as an "upper benchmark" (UB) representing how well the HBV-model could perform in the ideal case. To get a measure of how well the HBV model performs when there is no information at all about the catchment a set of 100 parameter sets were selected randomly, but still within the predefined parameter ranges, using a Monte Carlo procedure. The ensemble mean of the hydrographs produced by this process was named "lower benchmark" (LB). The upper and lower benchmarks were used to get a frame of reference for how well the other calibrations worked, as well as the index methods.

The remaining 4 ways HBV was calibrated were named (the bold word at the beginning of each paragraph) after what set of data they were based on, as this was the main difference between the calibrations:

1. **Waterlevel**: Calibration using daily water-level data from April 2016 to December 2020, using Spearman correlation as the objective function.

   1.1 **Waterlevel_MSE**: To investigate whether the water-level calibration could be improved by using the three available point observations of streamflow, a subset of parameter sets form *Waterlevel* was chosen. This was done by finding the ten parameter sets with the lowest MSE (Equation 4) when the discharge observations from 2018 were compared to the corresponding simulated value (thus no new calibration was done). As only three discharge observations existed these values were "extended" by assuming that every day with the same water-level (± 0.5 cm) also had the same flow. A new ensemble mean was created from the ten selected parameter sets.

2. **Point_Obs**: A calibration-method based on the three point observations of streamflow was carried out using a function of HBV-light. It uses limited discharge observations and, similarly to what was done for *Waterlevel_MSE*, searches in a file containing water-level data for all days with the same water-level as the observations (± 0.5 cm), and extends the observations to these days. The calibration period was the same as for *Waterlevel*, and the objective function was *ReffQObsSample*.

3. **Rating curve**: From the three existing discharge observations, and their corre-

sponding water-level, a rating curve was constructed for Håga river using a standard rating equation. As there were only three values to base the rating curve on the fitting was done visually, aided by knowledge of the stream (Appendix 8.4). Discharge values were calculated then for all days with a recorded water stage. The discharge time-series derived from the rating curve were used to calibrate HBV using the same calibration period as for *Waterlevel*, and KGE as the objective function.

4. **Donor catchment**: The final method was to calibrate the HBV model for one of the other basins, and then use the same parameter sets to simulate the streamflow in Håga. Using the results from the index methods Säva was chosen as the donor catchment. This calibration used data from the validation period (1991-2000), and KGE as the objective function.

As mentioned before the S-HYPE model was not calibrated specifically for this project. The data used to represent S-HYPE in further analyses was the daily simulated discharge supplied by SMHI.

3.6.3 Evaluation

All estimations of the streamflow in Håga river, DAR, MOVE, HBV-simulations and S-HYPE, were evaluated in the same way. The same evaluation (validation) period (1991-2000) was used for all methods, despite differences in the calibration periods. For each estimation every evaluation metric listed in Section 3.5.2 was calculated. Both daily and seasonal KGE-values were computed for the different methods as well. The seasonal evaluation was based on daily values divided into the seasons, where winter was defined as December to February, spring was March to May, summer June to August and autumn was defined as September to November.

For the HBV-simulations the evaluation metrics were only calculated for the ensemble mean of each calibration-method, but to get a grasp on the uncertainty of each calibration the KGE value of each simulation (100 per calibration-method) was calculated as well.

During the evaluation process the different metrics and signatures themselves were also compared to evaluate how useful they were to the analysis.

3.6.4 Hydrological changes over time

As the most recent validation data that exists for Håga is over two decades old it was decided to conduct three analyses that could serve as a foundation for the discussion on how relevant the different estimations are for the current hydrological conditions:

1. Change of land cover in Håga catchment. This was done by comparing orthophotos of Håga catchment. The photos that were taken closest in time to the evaluation period and the present were from 1973 and 2011 respectively, and were used as the basis of this analysis. Using these orthophotos Håga catchment was manually divided into areas of forest, open land, urban areas and open water, so changes between the two time periods could be spotted more easily.

2. Calibration using "old observations". As the *Point_ Obs* calibration of HBV only used discharge observations from recent years, it was considered of interest to see how the method would work if observations from the evaluation period were used instead.

To do this three values from the same hydrological year (1992) were picked from the discharge time-series, and an effort was made to pick values similar (in time and magnitude) to the modern point observations (Appendix 8.2). To use the intended calibration-method a water-level time-series was also required, which did not exist. Instead a synthetic water-level time-series was created from the discharge data, by replacing the discharge values with their respective ranks in the time-series, similarly to what was done by Pool et al. (2021). This way a data set containing information on the system dynamics, but without information on water volumes (analogously to water-level data), could be created. After the HBV had been calibrated using this data the result was analysed in the same way as the other HBV calibrations.

3. Changes in hydrological signatures for Säva catchment. As there are no current measurements of discharge in Håga the closest basin of similar size was used to study how some hydrological signatures have changed between the evaluation period and the present. The hydrological signatures used for this analysis were: MHQ, MLQ, FI, BFI, the runoff ratio (i.e. the amount of precipitation that becomes runoff) as well as the total amount of runoff and precipitation per year.

4 Results

The S-HYPE model is currently the only available source for present-day runoff estimates for the Håga catchment. Therefore, the simulated runoff from S-HYPE was used as a reference when comparing both the index methods and the HBV-simulations.

4.1 Index methods

The figure showing the performance of MOVE and DAR measured in KGE for the different donor catchments shows that the MOVE-method captures the streamflow of Håga river better than DAR (Figure 5). In all cases, except for Örsunda, the KGE-value is higher for MOVE (See Appendix 8.5 for exact values, as well as performance measured in Nash-Sutcliffe efficiency (NSE)). Säva is the best performing catchment and surpasses S-HYPE as well as the upper benchmark in KGE with the MOVE method. The worst results are from Sävja and Vattholma using DAR, which is worse than modelling with no information at all, and Vattholma is the worst performing catchment for both index methods.

When it comes to the KGE of the donor catchments for the different seasons Säva continues to be the best overall, although for spring and autumn Stabby performs better when using the MOVE-method (Appendix 8.5). Both methods and all basins generally struggle more to estimate streamflow during winter and summer.

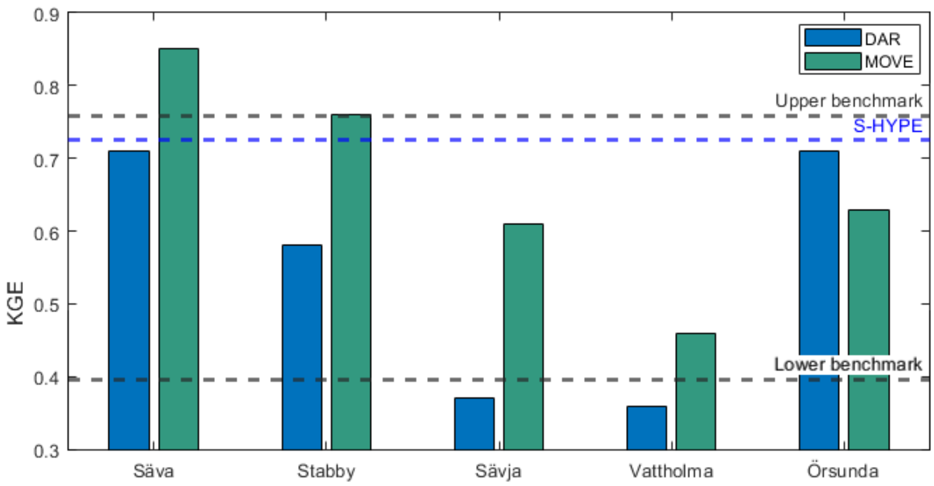

Figure 5. *The performance of the two index methods measured in KGE. The blue bar shows the performance of the DAR method, while the green bar shows the MOVE method. The lines represents different benchmarks used in the project.*

A comparison between the FDCs of the two different index methods shows that most basins, regardless of method, in general follow the same shape as the Håga river's FDC, but with a lower streamflow value (Figure 6). According to the figures MOVE performs better compared to the DAR method, as all curves are closer to the curve representing Håga river, in particular for the low flow-magnitudes. This reinforces that MOVE captures the streamflow in Håga better.

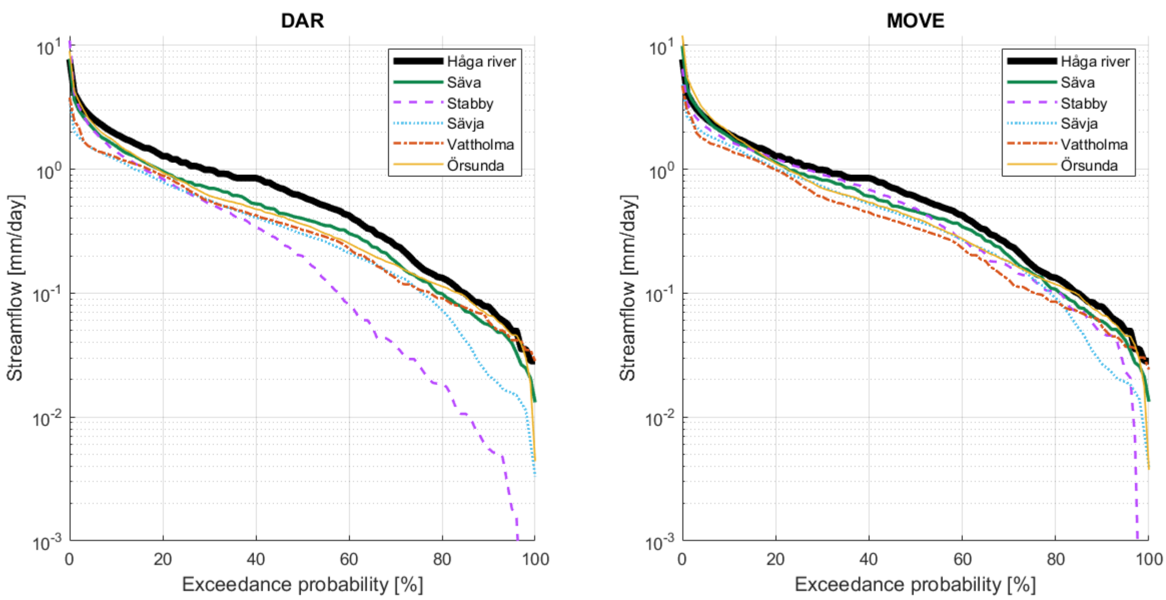

Figure 6. *Flow duration curves for the two index-methods. The thick black line is the actual FDC of Håga river for the evaluation period.*

The MOVE method using Säva as donor catchment continues to be the overall best approximation of Håga river, and surpasses S-HYPE both when it comes to capturing the dynamics of the system, and the overall water balance (Table 2). Vattholma seems to mirror the flow of Håga the least, and no matter method or donor catchment the amount of runoff estimated is always lower than the actual discharge of Håga river.

Table 2. Hydrological performance measures for the index methods. The closer to the "Håga"-value a method is, the better. The cells of the two best donor catchments for each method and signature are light green, while the best across both methods is a darker green. The worst donor catchment for each method is coloured red.

	Spearman	$Q_{diff}\,[mm/yr]$	FI	BFI
Håga	1	0	0.12	0.64
S-HYPE	0.91	-46	0.11	0.60
DAR				
Säva	0.96	-71	0.16	0.60
Stabby	0.92	-115	0.24	0.41
Sävja	0.92	-133	0.06	0.71
Vattholma	0.80	-120	0.04	0.90
Örsunda	0.89	-65	0.23	0.46
MOVE				
Säva	0.96	-29	0.16	0.59
Stabby	0.92	-46	0.15	0.59
Sävja	0.92	-82	0.07	0.71
Vattholma	0.80	-100	0.04	0.89
Örsunda	0.89	-18	0.25	0.43

Studying the hydrological signatures related to high- and low-flows (Table 3) confirms that Säva is the donor catchment that estimates the streamflow in Håga the best. Although it is not the best donor catchment for every signature, it is always at least second best, and always surpasses S-HYPE.

Table 3. Hydrological signatures related to high- and low-flows. "Freq." is short for "frequency of". The cells of the two best donor catchments for each method and signature are light green, while the best across both methods is a darker green. The worst donor catchment for each method is coloured red.

	MLQ $[mm/day]$	MHQ $[mm/day]$	Bias Q95	Bias Q90	Bias Q10	Bias Q5	Freq. lowflows $[d/yr]$	Freq. highflows $[d/yr]$
Håga	0.15	4.14	0	0	0	0	65	12
S-HYPE	0.07	2.82	0.22	-0.09	-0.29	-0.37	98	3
DAR								
Säva	0.10	4.20	0.13	-0.08	-0.20	-0.25	84	7
Stabby	0.03	5.35	-0.64	-0.77	-0.22	-0.23	162	8
Sävja	0.09	1.81	-0.11	-0.29	-0.50	-0.54	98	1
Vattholma	0.08	1.63	0.72	0.31	-0.57	-0.61	106	2
Örsunda	0.07	5.92	0.91	0.45	-0.15	-0.18	80	11
MOVE								
Säva	0.11	5.22	0.20	-0.02	-0.03	-0.08	81	13
Stabby	0.08	4.00	0.73	0.18	-0.24	-0.30	88	6
Sävja	0.11	2.38	0.11	-0.11	-0.35	-0.40	86	3
Vattholma	0.08	1.92	0.58	0.21	-0.50	-0.54	107	3
Örsunda	0.08	7.82	0.94	0.48	0.07	0.05	75	17

Unlike Säva, the other streams vary a lot more in how well their hydrological signatures match Håga: Säva is the only catchment that consistently ranked as the best or second-best donor (for the MOVE method).

The hydrograph below illustrates some of the earlier mentioned results: while the Vattholma estimation is consistently off the mark, Säva generally matches Håga well, while Stabby generally captures the high-flows better than the low-flows (Figure 7).

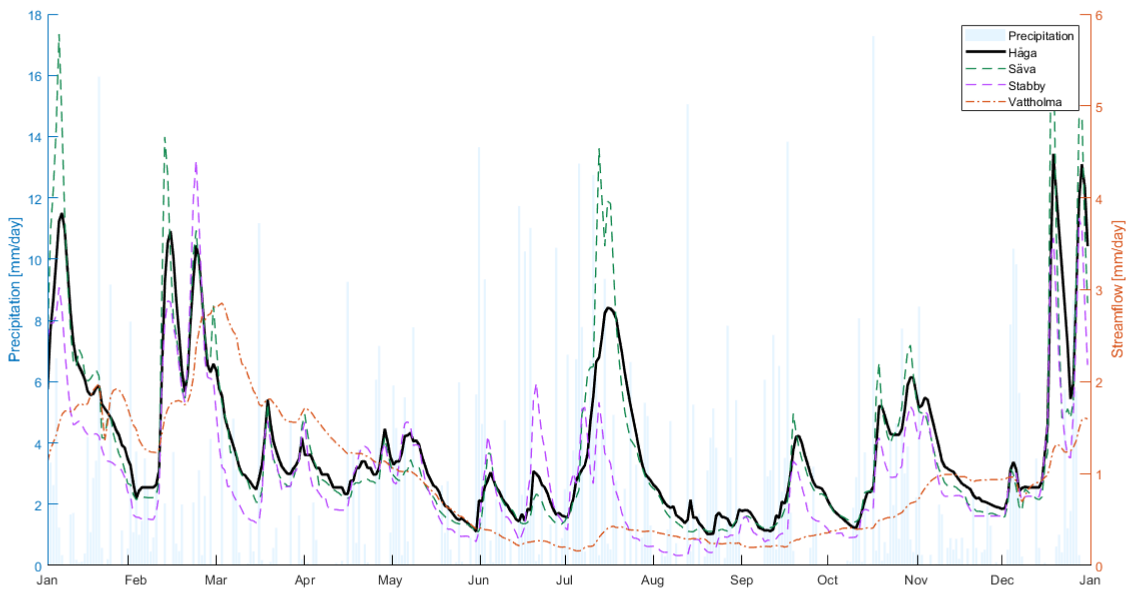

Figure 7. *Hydrographs for the year 1998 for some of the MOVE methods.*

4.2 HBV-simulations

The variability of KGE for the different calibration-methods differs a lot (Figure 8), but a general pattern is that the more data was used in the calibration, the smaller the variability of the method (the box-plot of UB and LB follows the same pattern (Appendix 8.6)).

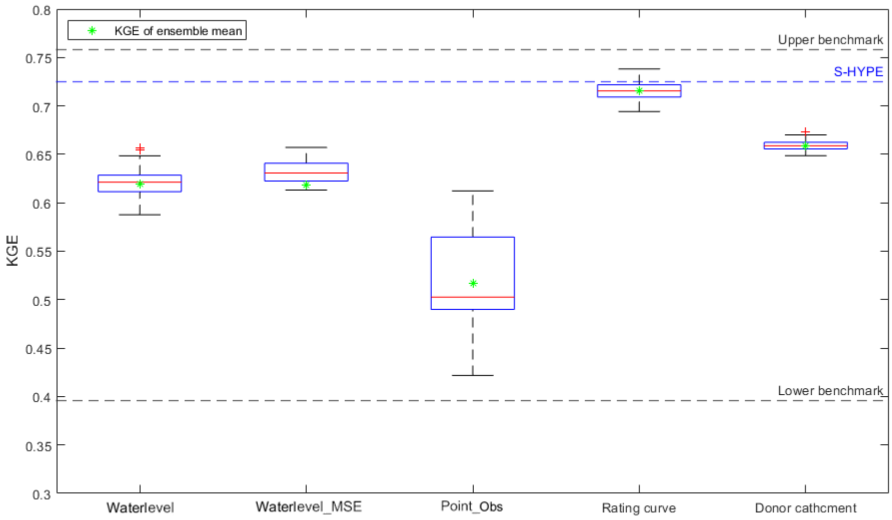

Figure 8. *The distribution of performance (measured in KGE) for the different HBV calibration-methods. All boxes are based on 100 simulations (100 values) except for Waterlevel_MSE, which is based on 10 simulations. The lines represents different benchmarks used in the project.*

Although the upper benchmark performs better than S-HYPE, no other ensemble mean surpasses the S-HYPE model when looking at KGE (Figure 8) (See Appendix 8.5 for exact values, as well as performance measured in NSE). The simulations from the *Rating curve* calibration are closest, followed by *Donor catchment*. Although UB does surpass the performance of S-HYPE, the difference is negligible. Concerning the two water-level-calibrations there is nearly no difference in the ensemble mean, but the KGE-values of *Waterlevel_MSE* are slightly higher than that of *Waterlevel*.

The KGE for the different seasons (Appendix 8.5) follow the same pattern as for the index methods where winter and summer have the lowest values. The KGE of the summer months were overall particularly low for the HBV-calibrations. The *Rating curve* calibration performs especially well during winter and spring, surpassing the seasonal KGE of UB. However, none of the HBV-calibrations surpasses S-HYPE for spring and summer.

When comparing the FDCs of the upper and lower benchmark, which represents HBV calibrated the ideal way (UB) as well as a situation where there is no data at all about the river (LB), it is apparent that LB has a very different behaviour with a less variable flow regime than Håga. UB's FDC is more similar to Håga's than S-HYPE, except when it comes to the lowest of flows (Figure 9). All calibrations of the HBV-model, as well as S-HYPE, seem to struggle most with the low-flow values; for flows with an exceedance probability of 90% or more, all FDCs, barring LB, are considerably lower than Håga's, and lower than S-HYPE's too (Figure 9). Similarly to the FDC's of the index methods (Figure 6), all curves are mainly below the curve representing Håga. Despite *Point_Obs* having a considerably lower KGE, its FDC is quite similar to the others, and even closer to HBV's curve for the lower flows (Figure 9).

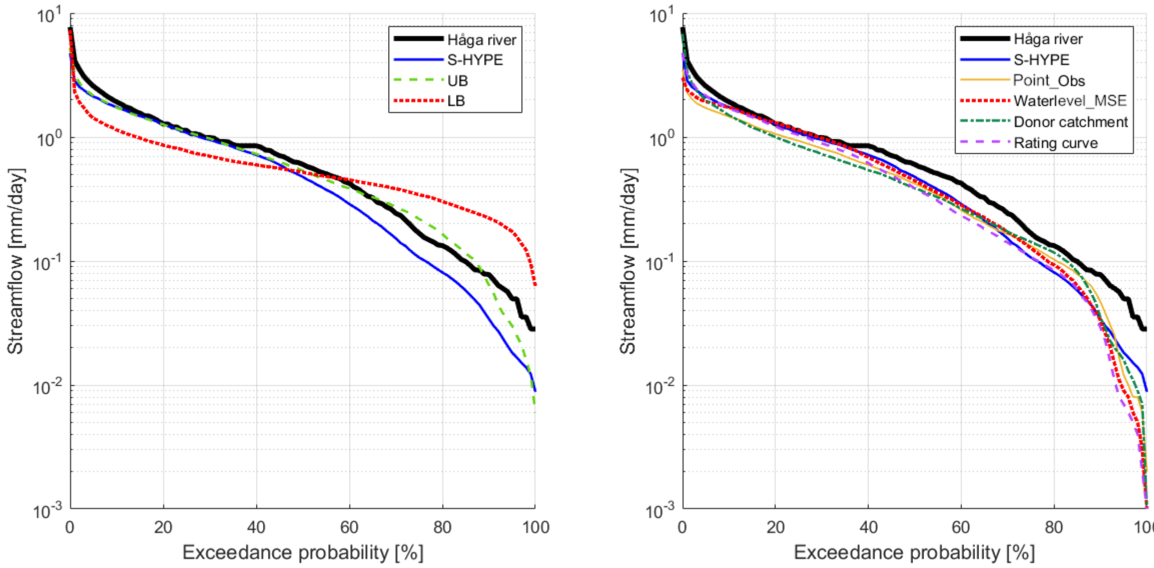

Figure 9. *Flow duration curves for HBV-simulations. To the left the FDC of UB and LB is shown together with Håga and S-HYPE as reference. To the right the references are show together with all calibration-methods except Waterlevel as it is essentially the same as Waterlevel_MSE.*

The value of the hydrological signatures can vary considerably depending on the calibration. However, there continues to be only a small difference between *Waterlevel* and *Waterlevel_MSE*, except for a slightly smaller error in the water balance for *Waterlevel_MSE* (Table 4 and 5). The water balance error of the HBV-simulations is, similarly to the index methods, always negative (Table 4), meaning that the total simulated runoff is too low. This is reflected in the HBV-calibrations tending to overestimate the number of low-flow days, while underestimating the frequency of high-flows.

Table 4. Hydrological performance measures for the different HBV-calibrations. The best calibration-method is coloured dark green, while the second best is light green. The worst method is coloured red.

	Spearman	$Q_{diff}\,[mm/yr]$	FI	BFI
Håga	1	0	0.12	0.64
S-HYPE	0.91	-46	0.11	0.60
UB	0.90	-28	0.10	0.61
LB	0.76	-74	0.13	0.69
Waterlevel	0.86	-59	0.06	0.75
Waterlevel_MSE	0.86	-54	0.06	0.75
Point_Obs	0.86	-86	0.06	0.75
Rating curve	0.89	-58	0.09	0.61
Donor catchment	0.90	-76	0.15	0.52

Despite UB being supplied with the most informative data, it does not always matchs the hydrological signatures the best, although it is consistently "good" compared to the other calibration-methods. In most cases UB is better than, or on par with, S-HYPE, except for when it comes to bias for the low-flows (Q95 and Q90) (Table 5).

Despite *Rating curve* having the second-best KGE-value after UB (Figure 8), it is rarely one of the best calibration-methods for any hydrological signatures. However, it is never the worst and is consistently comparable to S-HYPE (Table 4 and 5).

Table 5. Hydrological signatures related to high and low-flows for the HBV-calibration-methods. "Freq." is short for "frequency of". The best calibration method is coloured dark green, while the second best is light green. The worst method is coloured red.

	MLQ [mm/day]	MHQ [mm/day]	Bias Q95	Bias Q90	Bias Q10	Bias Q5	Freq. lowflows [d/yr]	Freq. highflows [d/yr]
Håga	0.15	4.14	0	0	0	0	65	12
S-HYPE	0.07	2.82	0.22	-0.09	-0.29	-0.37	98	3
UB	0.10	3.33	0.87	0.57	-0.25	-0.31	57	5
LB	0.18	3.13	4.30	3.51	-0.53	-0.57	7	2
Waterlevel	0.08	2.14	1.23	0.44	-0.39	-0.47	90	0
Waterlevel_MSE	0.08	2.15	1.44	0.57	-0.38	-0.46	89	0
Point_Obs	0.08	2.12	0.45	0.06	-0.45	-0.51	84	1
Rating curve	0.07	3.07	0.00	-0.25	-0.27	-0.32	99	5
Donor catchment	0.06	3.82	-0.14	-0.19	-0.29	-0.31	76	6

LB is the "calibration" that approximates the hydrological signatures worst most of the time, but despite not using any actual information from Håga it manages to approximate some hydrological signatures well. The observation-based calibration is in many cases the worst after LB.

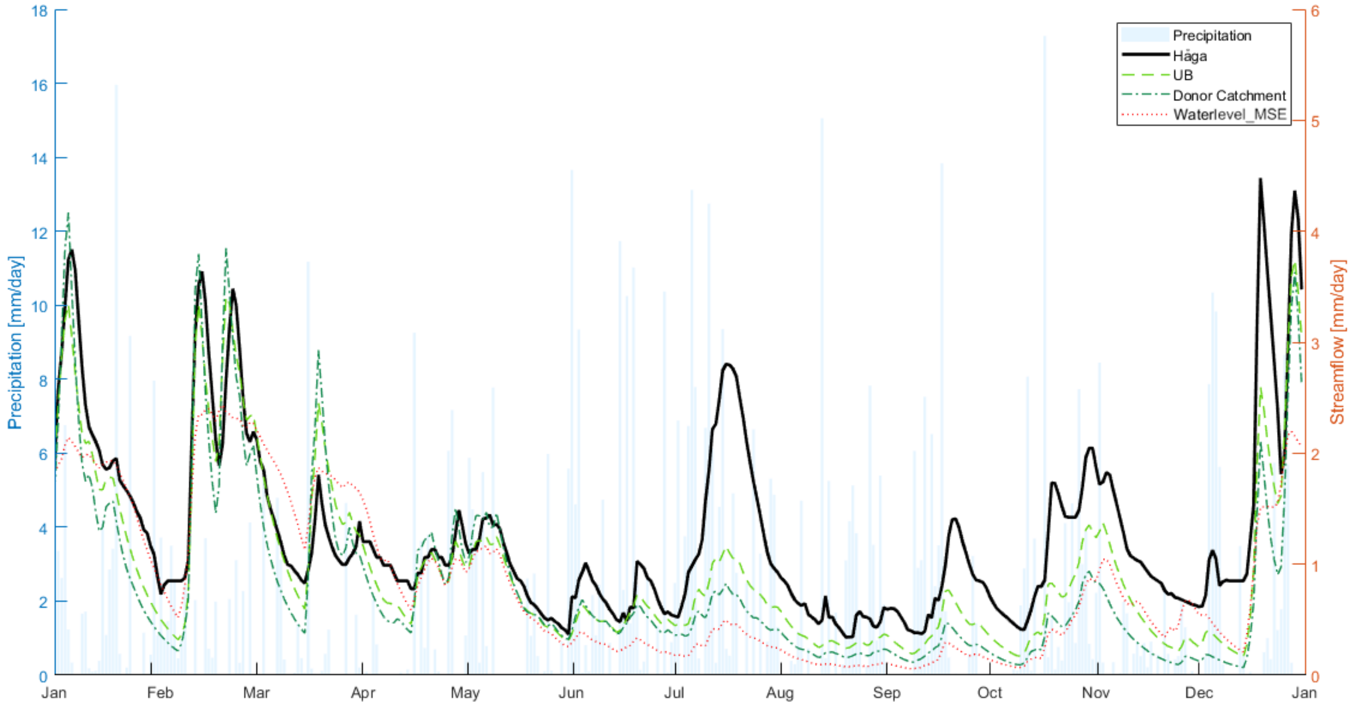

Figure 10. *Hydrographs for the year 1998 for some of the HBV-calibration-methods.*

The hydrograph (Figure 10) illustrates some of the results presented in the tables. For example it is apparent that *Waterlevel_MSE* has no "high-flow days", as its highest flows are just above 2 mm/day. While UB and *Donor catchment* are quite similar, the hydrograph of *Donor catchment* is consistently below UB, which is reflected in the larger Q_{diff}. The graph also shows that all simulations struggle more during summer (Figure 10).

4.3 Hydrological changes over time

Between 1973 and 2011 the land use of Håga catchment has changed, and forest covers larger parts of the area in the later period. In addition, the amount of urban area has also increased, particularly near the outlet of the basin, close to central Uppsala.

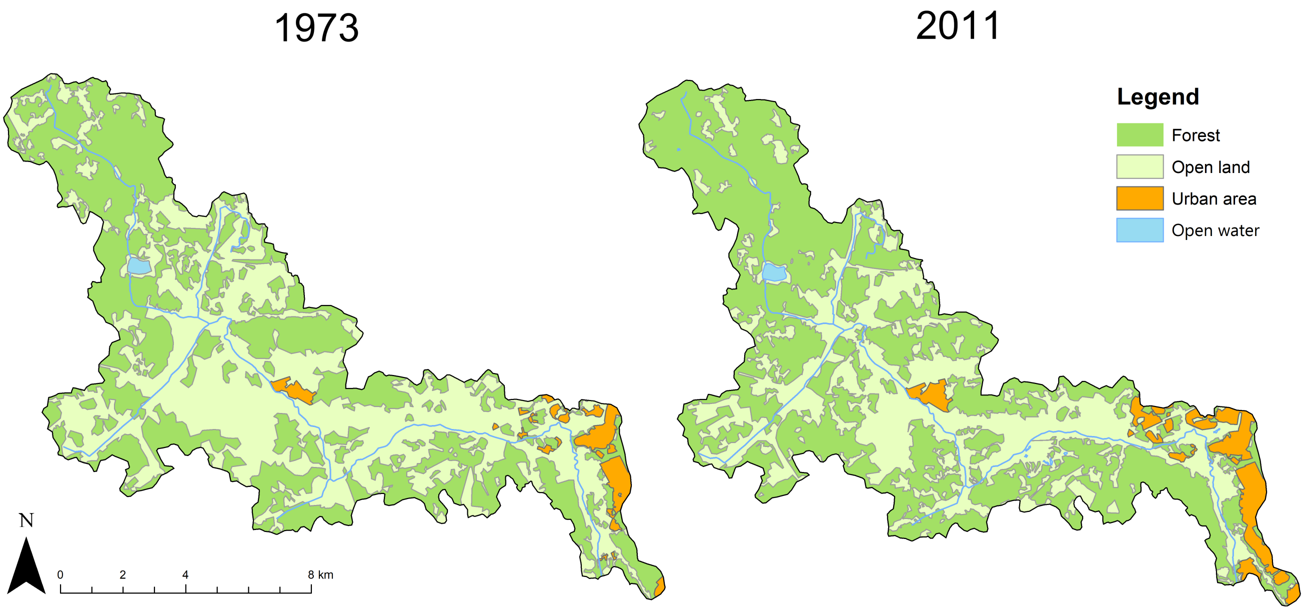

Figure 11. *Ground cover changes in Håga catchment between 1973-2011. The category open land includes both meadows, arable land, and clear-cut areas. The Håga river starts in the north-western part of the basin and has its outlet in the south-eastern part.*

The calibration of HBV using values from 1992 yielded considerably higher KGE-values than the one using data from 2018 (Figure 12). Not only are the KGE-values higher, but the calibration using the older data set performs better for every single hydrological signature, barring bias for Q90 and Q95 (Appendix 8.7). However, the artificial water-level time-series used in the calibration with the old point observations was less precise than the actual water-level data from the gauge in Håga. This resulted in more than twice as many days when the stream level was the same as one of the point observations (47 for the modern period, 117 for the old), and thus also resulting in that many more data points being used in the calibration with the older data.

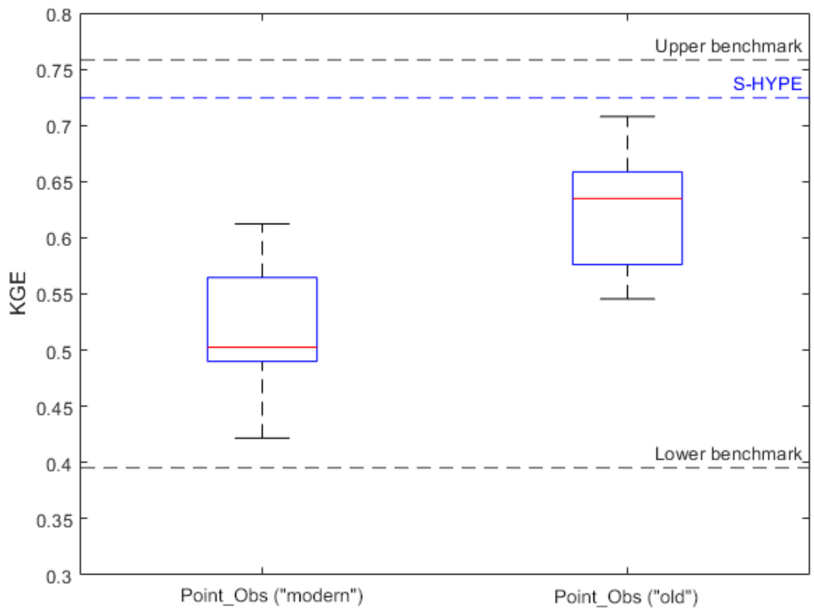

Figure 12. *The distribution of KGE-values for HBV-calibrations using either more recent point observations of streamflow or "synthetic" observations from the evaluation period. Both boxes are based on 100 values*

Comparing the hydrological conditions of the river in Säva catchment reveals that there is overall less runoff in the more recent years (Table 6). The runoff ratio of the basin has also been reduced, indicating that a smaller portion of the precipitation becomes streamflow. The hydrological signatures that are not directly related to water volumes, FI and BFI, do not seem to have changed as much.

Table 6. Hydrological signatures for Säva catchment during the periods 1980-1990 and 2010-2020

Period	MLQ [mm/day]	MHQ [mm/day]	Q_{tot} [mm]	P_{tot} [mm]	Q/P	FI	BFI
1980-1990	0.10	6.00	291	661	0.44	0.16	0.54
2010-2020	0.03	4.57	208	603	0.35	0.16	0.51

5 Discussion

The discussion is divided into sections representing the different research questions and the research questions are discussed separately and in the order they were presented to the greatest extent possible (Section 1.1)

5.1 Index methods

Regionalising streamflow from other catchments to Håga can work very well: one of the index methods even produces the overall best performing estimation of all methods used, including the hydrological models. In most cases the index methods performed better than the randomised parameter sets (LB). For example all methods, regardless of the catchment (except for Stabby using DAR), produced FDC:s that follow the general shape of Håga, in contrast to LB's which is completely different. However, the performances of the index methods vary considerably depending on donor catchment and whether DAR or MOVE was used.

The fact that one of the index methods could out-perform the hydrological models should however probably more be seen as Säva being a remarkably good donor catchment for Håga, rather than a general result pointing to index methods being superior to models. A study by Marahatta et al. (2021) comparing multiple index methods, including DAR, with a hydrological model (SWAT) for a Nepalese catchment, actually found the opposite, suggesting that using a suitable hydrological model is the better option for streamflow estimations at a monthly time scale.

However, adjusting a model to a catchment requires more effort than index methods like DAR. Seibert (1999) found, when studying catchments in the Uppsala area using HBV, that an ensemble mean of specific runoff from the streams in a region (i.e. a regional DAR time series) performed almost as well as HBV after it had been adapted to the region. Creating an ensemble mean of DAR also decreases the risk of very bad predictions, making it a safer option than using DAR with one donor catchment. So although index methods should not be considered superior to hydrological models, they can be a good alternative in the cases where only a runoff time series is required.

Out of the two index methods investigated in this project MOVE is the one that performs better, almost always producing both higher KGE-values and closer matching hydrological signatures than DAR. The fact that MOVE produces better results than DAR is not that surprising, considering that it uses data from Håga itself. MOVE is also *always* better than LB KGE-wise, and for many hydrological signatures as well.

The different donor-basins do have different strengths, as some of them only perform well on certain aspects of the flow. For example Stabby does quite well at estimating the high-flows of Håga, as well as the conditions during spring and autumn, while Sävja does quite well with low-flows. All this is however overshadowed by Säva, which by far is the best donor catchment for the methods used in this project. When using MOVE Säva always produces one of the best two estimations for all the hydrological signatures, and as a result also has the highest KGE. In addition to this it almost always performs better than S-HYPE.

Something both index methods, as well as the hydrological models, have in common is that they underestimate the amount of total runoff in Håga, in most cases estimating both too many low-flow days and too few days with high flow. The question is if this underestimation of volumes is due to some "special" conditions surrounding Håga, or connected to the estimation methods themselves. Some possible explanations could be the high runoff ratio in Håga, the fact that the runoff ratios decrease for all other catchments (Appendix 8.1) or uncertainties in the catchment areas (for DAR). Another possible explanation is simply that there are errors in the streamflow data for Håga, and that the discharge values are higher than they should be. This could have happened for example if some debris caused a dam just downstream of the measuring station as that would lead to a higher water level, which is usually what is measured.

One other thing both index methods have in common is that they capture the timing of flow events well, and thus all have quite a high Spearman value; many being better or comparable to S-HYPE and the best of the HBV-simulations. This was rather surprising as the index methods do not change the timing from the donor catchments, and were thus expected to do worse than the hydrological models that are adapted to Håga and use meteorological data from that specific area. Nevertheless, the index methods did have overall high Spearman values, which might have to do with the catchments being similar and close enough to have almost the same timing of flow peaks.

The actual importance of estimating the timing of flow peaks is however not always important: if a flood happens one day or two days later probably does not matter when estimating flood risk. In this aspect using nearby basins as donor catchments might not be as important: finding catchments that estimate the magnitude of flows well should be the focus instead. In these cases it is also important to be mindful of what goodness-of-fit functions or hydrological signatures are used, and not use ones like Spearman and KGE that are based on correlation and thus get low values if the peaks do not match.

5.1.1 Preferable donor catchment

Based on how much better Säva works as a donor catchment than the other basins it is clear that the choice of donor catchment matters. From just studying five different donor catchments and six traits it is nevertheless quite hard to draw a clear conclusion on why a certain basin works better, and what catchment trait matters the most.

Although Säva is is most similar to Håga in regards to catchment area and areal percentage of lakes, which argues for the importance of these traits. The lake-percentage in particular seems important as Vattholma, the overall worst donor catchment, also is most different from Håga in this regard. The position of the lakes probably also plays a role: in both Håga and Säva the lake in question is located high up and on the outskirts of the catchment, while in Vattholma the lakes are closer to the gauging station in the catchment.

It seems reasonable that both lake-percentage and catchment area would be important. Lakes are known to sustain baseflows as well as attenuate and delay flow peaks (Leach et al. 2019); theories that are reflected by Vattholma's low FI and Spearman value, as well as its high BFI. The area makes a difference as bigger catchments both have longer flow paths and tend to be more ground-water dominated (Blöschl et al. 2013). This way the catchment size can matter even though specific discharge is studied, something seen by the fact that Sävja too has a high BFI.

Another factor that works in favour of Säva is probably the fact that it is located so close to Håga. A study by Pool et al. (2019) found that spatial proximity worked better than trait-similarity for finding good donor catchments. In this case all catchments are located quite near each other, and Säva also happens to be rather similar to Håga trait-wise, but the fact the two basins are side by side probably still makes a difference in itself. This can be seen by the high Spearman value for Säva, which is probably because rain frequently falls at the same time in the two catchments. It is of course also possible that Säva is similar to Håga when it comes to some catchment trait that was not analysed, such as soil types.

5.2 Hydrological models

Neither S-HYPE nor HBV performed better than the best performing index method. There are however many other uses of hydrological models than just recreating flow records, making these methods important nonetheless.

5.2.1 Comparison between S-HYPE and HBV

The model fits of HBV in this project are along the lines of the performance found in other studies of rivers in the Uppsala area using HBV (Seibert 1999). This comparison is however a bit unfair as the other study both calibrated against, and measured the model performance in, NSE. Although the performances have been calculated in NSE for the calibrations of this study as well (Appendix 8.5) the NSE-values would likely have been higher for the calibrations if NSE was used as the objective function.

When it comes to S-HYPE the performance of the model varies a lot depending on the catchment and is generally lower for small ($<$200 km^2) catchments like Håga (SMHI 2022c). The NSE-values vary a lot between 0.3 to 0.85, with the performance in most small catchments in being between 0.5 to 0.7. S-HYPES performance in Håga is thus on the higher end compared to other catchments in Sweden. This might have to do with the river in Säva catchment, as it is used for calibration of S-HYPE: something that would benefit Håga as the catchments are so similar.

Although the calibrations of HBV in this project vary a lot in performance, it is possible to produce better estimations with HBV than what is supplied by S-HYPE, despite HBV being a simpler model and S-HYPE performing on the high end of the spectrum. Though for this to be the case HBV has to be calibrated the ideal way (UB), and even then there is barely any difference between the KGE of the two models. UB still manages to be better or equal to S-HYPE for almost all hydrological signatures, and is considerably better at estimating the frequency of low-flows.

Even when not considering UB there is always some other calibration-method that is on par with, or better than, S-HYPE for each of the hydrological signatures. S-HYPE is also out-performed by several of the calibration-methods when it comes to MLQ, MHQ and the frequency-related signatures. Despite being calibrated with both more limited and more uncertain data than UB, the *Rating curve* calibration is also often on par with S-HYPE.

Overall, both S-HYPE and HBV struggle with capturing the high- and low-flows: both models overestimate the number of low-flow days and underestimate the frequency of high-flows, and also underestimate the total amount of runoff in Håga. The reason the two models have troubles in the same areas could be because S-HYPE is based on HBV (SMHI 2020), and since extremes simply are harder to model. This is in agreement with other studies, for example one by Van Kempen et al. (2021) who also found that the timing and magnitude of extreme flow events in cold and temperate zones were more affected by the model structure and parameter sets than the simulations of arid and tropical areas were. Why the total runoff is underestimated is hard to say, but precipitation data and evaporation fluxes might be of importance, or Håga's comparably high runoff ratio, as mentioned before.

A somewhat unexpected result was that, despite using regionalised parameters, S-HYPE captures the timing of flow peaks in Håga slightly better than HBV, which uses parameters specifically calibrated for the catchment of interest. Maybe this is due to the precipitation data that is used as input, in combination with the regionalised parameters working quite well.

Another area where S-HYPE performs well is during spring and autumn, where it is on par with the best performing index method, while it has a lower performance during summer and winter. The variation between the performance of the different seasons is bigger for S-HYPE than HBV. The reason for this is hard to say, but is probably due to a difference in how the two models work and how S-HYPE is calibrated.

This is another difference between S-HYPE and HBV; there is not a lot of information on how the HYPE model is calibrated for Sweden, and how the calibrations differ for different areas and times. For example there was an SMHI stream gauge in Håga during the period the S-HYPE values were taken from, but it is unclear whether these measurements themselves were used for the calibration or not. In this sense HBV is favourable, as there is more understanding of the input data and the calibration process at the same time as it needs less data to be used, and performs on a level comparable to S-HYPE. Other studies have also found that the performance of HBV and S-HYPE are comparable (Gustavsson 2014; Reynolds et al. 2021), and one study has, in particular, noted that the difference in data requirements to run the two models is considerably larger than the difference in performance (Gustavsson 2014).

5.2.2 The effect of differing calibration data

The calibration-methods and the data used matter a lot for how well the HBV-model works, both for the overall performance, and the different hydrological signatures. However, no matter the amount or type of data the calibration-methods always worked better than the simulations that were randomised and used no data (LB). This is in agreement with the study by Seibert et al. (2009) who only found decreases in performance when two discharge measurements or fewer were used for calibration, while using more data would increase the performance. Only using limited or alternative data did not provide as good fits as a "full" calibration based on streamflow data, but these calibration attempts show that even alternative data is useful when modelling ungauged basins in general. It is however important to note that the quality of the data plays an important role too, and that the data needs to be representative to be useful. An example can be seen in the difference in performance between the *Point_ Obs* calibrations using old and new data, something that likely is related to a difference in how representative the data is, and its quality.

Not only the type of data matters, but also the amount of data. From the box plots (Figure 8) it is apparent that the calibrations based on more data also are less variable.

The objective function used for calibration could also matter in this case, as all calibration-methods using KGE also had a smaller variability in KGE. On this note comparing the different calibration-methods fairly can be quite difficult, as both the objective function and the amount of input data vary for the different methods.

Something general for both the index methods and the hydrological models is the fact that estimations work better during spring and autumn. A recent study by Pechlivanidis et al. (2020), using the HYPE model for Europe, found a different pattern as seasonal predictability was highest during spring and summer. In this study they reasoned that it could be related to smaller discharge volumes in the warm months. However, as KGE (used to calibrate several of the HBV-models) is known to prioritise high-flows, it is not odd that these models do not work particularly well for low-flows, and in extension during summer. This does however not explain the lower predictability during winter, or why the pattern is the same for the models using other objective functions (nor that the same pattern is found for the index methods). Instead, an explanation might come from conditions surrounding snow and drought, or other extremes, that are more common during winter and summer. During winter more uncertain validation data, due to ice disturbing the measurements, might also play a role.

It is important to note that the performance of HBV depends a lot on how the model is calibrated, and that every single hydrological signature could be captured better if an objective function that prioritised that signature was used, or if the signature itself was used in the calibration. In this way the objective function is very important, which also is seen in the results. For example the calibrations using KGE, which somewhat prioritises high-flows, captures the frequency of high-flows better than the other calibration-methods.

One rather unexpected finding of this project concerns the KGE-values of the ensemble means. Other studies have found that the ensemble mean consistently performs better than the average individual parameter set. This is because making ensembles tends to smooth out errors (Seibert 2022; Seibert et al. 2009), which was one of the reasons they were used. This was however not the case in this study, where the KGE-values of the ensemble means always were found in about the middle of the boxes.

When it comes to each individual calibration-methods they all have their own strengths, weaknesses and interesting aspects. For example, despite being calibrated against 10 years of discharge data, UB still only has half as many high-flow days as it "should". This again probably has to do with KGE's tendency to underestimate peak flows. When it comes to LB it does capture FI, BFI, MHQ, and MLQ as good or better than many of the other estimations. This is however most likely just a "lucky coincidence" and should not be seen as LB capturing the system, as it is apparent that it does not from the FDC.

Considering that the water-level-calibrations (1) contained no information on water volumes and (2) the data was recorded around two decades after the validation period, the simulations performed well; even beating calibration-methods that used data containing water volumes when it came to Q_{diff}. This is in line with what was found by Seibert et al. (2016): simulations using Spearman rank as objective function can simulate streamflow well. The calibration-method does have obvious weaknesses however, for example it does not produce any high-flow days at all. Despite having Spearman as objective function it does not have particularly high values for Spearman rank either. This is probably in part due to the time gap between the calibration and validation period. However, as the

Rating curve calibration, which uses calibration data from the same period as *Waterlevel*, has a higher Spearman-value this cannot be the only explanation. Perhaps having information pertaining to water volumes helps in capturing the stream dynamics for the HBV-model.

Combining the water-level data with point observations of streamflow only marginally improved the performance. The fact that only three observations were available probably contributes to the fact that the difference was so small. However, as *Waterlevel_ MSE* was created from a subset of *Waterlevel* parameter sets, it is also possible that there simply were no simulations that fit the point observations well to choose from. After all, there were "only" 100 streamflow simulations, and since water volume is not considered in the water-level-calibration the actual volumes can vary a lot from simulation to simulation (Seibert 2022). It is entirely possible that there simply did not exist enough, or any, simulations that matched the point observations well. Using the observations to pick the best parameter sets would in that case not add much additional information.

When it comes to the calibrations based on water-level data some uncertainties have to be considered. Firstly, the data has several, up to a month-long, gaps and is thus incomplete. Secondly, the stream level dynamics captured by the gauge do not always represent the streamflow dynamics. One example is when a tree falls into the river and raises the water-level without decreasing the flow: something that did happen in 2020 and could happen again.

The *Point_ Obs*-calibration seemed to be the worst calibration-method at first. However, when the same method was used in combination with older data the result was much better: the method performed as well as *Waterlevel_ MSE* KGE-wise, but captured BFI and FI better. The poor performance of the calibration based on point observations thus likely has a lot to do with the data, which represented Håga inadequately either due to the time gap between calibration and validation period, or because the point observations were from 2018, which was an unusually warm and dry year in Sweden (Sjökvist et al. 2019), and thus not representative. Nonetheless, it is apparent that this type of calibration was the most uncertain, undoubtedly because they were based on only three streamflow observations, which is not enough information to constrain the model parameters despite these observations being extended to other days with the same water-level.

Concerning the *Rating curve* calibration, it is surprising that it performed so well; it is both based on modern data and only 4 years of it, in comparison with UB which has 10 years of data much closer in time to the validation period. In addition to this, the rating equation was only based on three streamflow measurements, which should make it very uncertain even though its construction was aided by knowledge of the stream. The reason it produces such good simulations is thus rather unclear, but it might have to do with the three measurements covering a large range of flows, or maybe the 4 years used for the calibration were particularly informative. Something that could have been done to investigate the uncertainty of the rating curve closer is to have intentionally introduced errors to see how they affected the performance (something that could have been done for all data sets used for calibration). Nonetheless, in addition to the uncertainty of the rating equation itself, this calibration-method also suffers from the same uncertainties as the water-level calibration. One must also keep in mind that the rating equation of a stream can change with time (Blöschl et al. 2013).

Despite performing worse than the *Rating curve* calibration, the *Donor catchment* calibration is the one with the least variation in KGE. This is most likely because it both uses KGE as objective function and the largest amount of data (10 years). The *Donor catchment* calibration does work well, which is expected as Säva worked well as a donor catchment for the index methods. However, some things seem to be lost when using HBV as an in-between step, as the overall performance is worse than using DAR with Säva as donor catchment.

5.3 Hydrological changes over time

One major uncertainty in the project is the fact that the most recent validation data is over 20 years old. This brings to question how relevant the results are for the current hydrological conditions, which in turn depends on how much the hydrology has changed.

Something that could point toward a change in hydrology is the change in the land use in Håga; many areas that were open land have become forests, and urban areas have grown - increasing the number of impermeable areas and consequently reducing infiltration. Changes in land use can alter the runoff dynamics of a catchment significantly, and specifically an increase in woodland areas is known to lead to an increase of evapotranspiration and thus that a smaller fraction of rain becomes runoff (Bosch et al. 1982).

Just based on how the land cover has changed it seems reasonable to assume a change of hydrology, and the differences in the HBV calibrations based on old and new point observations of streamflow support this. For example the calibration based on modern data resulted in simulations with less total runoff, which could be reflecting the increase of forested areas. It is however likely that the difference in the simulation was affected by other things than a change in hydrology itself. As mentioned before the modern streamflow point observations were taken from a year that can be considered unusual, and probably had lower flows than normal, which also could be the reason for the smaller amount of total discharge. Beyond this, the fact that the old water-level time-series was artificial also matters as it turned out to be less precise than the measured water-level time-series. It thus led to more days when the stream level was the same, increasing the extension of the three point observations. The method using the old data thus had more data points for the calibration, which might be a reason for its higher performance, rather than the old point observations actually representing the period better.

Considering all the uncertainties surrounding the observation calibrations it is hard to draw certain conclusions. The best alternative would be a comparison of streamflow data from the evaluation period and the last ten years from Håga, but as this was not available Säva catchment was studied instead, as it is located close to Håga and could have gone through similar changes. From the comparison it is apparent that there is less runoff during the modern period. This is of course linked to the smaller amount of precipitation, but the change in runoff ratio indicates that it is also due to a smaller fraction of the rainfall turning into discharge. As runoff ratio is connected to evapotranspiration as well as water loss to deep groundwater (McMillan 2020), a warmer and drier climate could explain the change. However, it is also possible that Säva has undergone a similar change in land cover as Håga, and that an increase in forested areas could be contributing.

When it comes to FI and BFI, the hydrological signatures that are less directly related to volumes of water, the changes are non-existent or small, indicating that the catchment continues to work similarly, despite smaller amounts of water. As the changes in water volumes probably are related to circumstances like climate, which would be the same for the two catchments, it is likely that Säva still works well as a donor catchment for Håga for the index methods. This is also supported by the fact that the MOVE-method have been used to extend streamflow records for 50 years (Hirsch 1982).

It seems like most changes likely to have happened in Håga would be related to streamflow volumes decreasing. However, other changes are not unreasonable either due to the increase of urban area which might have resulted in some business that for example extracts water for irrigation.

What effect these changes have on the performance of HBV is hard to say, but even if the performance decreases it is possible it still is in line with, or better than, S-HYPE (as it is not known how its performance has changed either), but with the added benefit of being in control of the calibration.

5.4 The usefulness of evaluation metrics

During this project many different metrics for evaluating performance have been used: hydrological signatures as well as goodness-of-fit functions. Although an effort was made to consider all metrics equally, some of the signatures ended up being more useful.

Even though an initial selection was made it was apparent, as discussed by Olden et al. (2003) and many others, that most of the hydrological signatures were correlated, especially those related to high- and low-flows. Although no formal correlation analysis was executed it was easy to find patterns: methods with a low MLQ also often had a higher frequency of low-flow days, a smaller bias for Q90 and Q95 as well as a FDC that had lower streamflow values for a given exceedance probability than other FDCs. Similarly, a low FI often meant a higher BFI. Thus it would probably not have been necessary to use all of the hydrological signatures, although they in this case had a purpose in helping reveal the usefulness of other signatures.

One signature I believed would be informative but turned out to be somewhat useless was the bias signatures. First and foremost they were quite hard to understand: although the sign of the number indicated whether there was too much or too little water, the magnitude was harder to grasp, and it did not seem fair to compare a "0.3" for Q95 with the same value for Q5. It was also impossible to tell whether it was the timing or just the water volume that was wrong. Apart from being hard to grasp it could also be directly misleading if studied on its own. For example, many of the methods with a Q90- or Q95-bias particularly close to 0 often had a way too high frequency of low-flow days. At the same time the methods that had a more reasonable low-flow frequency sometimes would have a much higher bias, probably just because they had a lower probability of matching the right days when they had fewer low-flow days to begin with.

In a similar vein, the signatures were often perceived as more useful if they were easy to understand, such as the FDCs, Q_{diff} as well as the high- and low-flow frequency values. Although the FDCs were easy to understand they were not always all that informative, for example the *Point_ Obs*-FDC looks about as good as those of the other methods (Figure 9), while it has a much lower performance for almost all hydrological signatures. MLQ and MHQ were also easy to understand, but due to them only being based on one value per year they were seen as rather uncertain. Nonetheless, they still contributed by giving an understanding of flow magnitudes. BFI and FI were interesting signatures as they provided something completely different than the volume related signatures, but they were also a bit difficult to interpret, especially since I had no reference for what counted as a "big difference", and what that would mean in reality.

The overall assessment using KGE felt very useful in general, as methods with a high KGE often performed well for most hydrological signatures. Just using KGE on its own did however not feel like a fair option, at least not in the case of the hydrological models. This was because some methods used KGE as their objective function while others did not, and the ones who did also had the highest KGE values.

6 Conclusions

The overall best, and also quite simple, way to estimate the streamflow of Håga river is using the index method MOVE using data from the donor catchment Säva. However, in a case where there is no streamflow data using the index method DAR is very uncertain, as the performance varies tremendously between donor catchments. A possibly better idea in this case could be to look at an ensemble mean of several nearby catchments. When it comes to the traits of the donor catchments in question it seems like the amount and placement of lakes as well as the size and location of the catchment itself matters most, and Håga is probably quite "lucky" to have a donor catchment that works as well as Säva, as this most likely is a big reason the MOVE method could out-perform the models.

While index methods excel at recreating streamflow records, a model has considerably more uses. In a case where a hydrological model is needed HBV seems like a better choice in comparison to S-HYPE, as it is both a simpler model but also allows more control and insight into the calibration, while performing equally well. The data used for calibration matters a lot though, and to decrease uncertainty it is important to use as much data as available. However, all data sets used in this project increase the performance of HBV in comparison to a case without data, and combining different types of limited or alternative data is superior to using a single data type, even though the difference is small. Although reaching the performance HBV has when it is ideally calibrated is difficult, just having some data can go a long way in catchments where there is no continuous streamflow record. It is however important that the data is of good quality.

Even though it is apparent that the hydrology of Håga catchment has changed, it seems likely that Säva still would remain a good donor catchment. However, there are many uncertainties surrounding this, which is something that could be looked further into in a future Master's thesis. For example the dynamics of the current water-level measurements in relation to rainfall could be compared to the dynamics during the '80s and '90s, to see whether the lag time between rainfall and runoff peaks has changed. It would also be interesting to do a similar study in Säva, as well as map the change of land cover for Säva similar to what was done for Håga in this project. Something else that would be interesting is to look further into the question of why all methods underestimated the flow in Håga - *is* there something special with this basin in particular? This investigation could start by examining the runoff ratio of Håga and nearby catchments in relation to land use and other relevant catchment traits.

When it comes to the different evaluation metrics used in this project it is evident that many are correlated. Although some probably were redundant they contributed by showing different aspects of the same phenomena. If nothing else, all of them were useful for gaining more insight about hydrological signatures in general, and in particular by revealing the weakness of other metrics.

Models, methods, and metrics aside: if a truly good record of streamflow is desired for Håga the best idea would probably be to finish the work of creating a rating curve for the river. There are certainly difficulties surrounding the risk of debris build-up in the stream, and the fact that the rating equation can change over time, but all other streamflow estimation methods also come with their uncertainties. Since a water-level gauge already is installed it would be reasonable to use it to its full extent, as it seems unlikely that any estimation methods could beat actual measurements in the river.

7 References

References

Addor, N., Nearing, G., Prieto, C., Newman, A., Le Vine, N., and Clark, M. P. (2018). "A ranking of hydrological signatures based on their predictability in space". In: *Water Resources Research* 54.11, pp. 8792–8812.

Atkinson, S., Woods, R., and Sivapalan, M. (2002). "Climate and landscape controls on water balance model complexity over changing timescales". In: *Water Resources Research* 38.12, pp. 50–1.

Baker, D. B., Richards, R. P., Loftus, T. T., and Kramer, J. W. (2004). "A new flashiness index: Characteristics and applications to midwestern rivers and streams 1". In: *JAWRA Journal of the American Water Resources Association* 40.2, pp. 503–522.

Bárdossy, A. (2007). "Calibration of hydrological model parameters for ungauged catchments". In: *Hydrology and Earth System Sciences* 11.2, pp. 703–710.

Bergström, S. (1990). *Parametervärden för HBV-modellen i Sverige: erfarenheter från modellkalibreringar under perioden 1975-1989*. SMHI.

Blöschl, G., Bloschl, G., Sivapalan, M., Wagener, T., Savenije, H., and Viglione, A. (2013). *Runoff prediction in ungauged basins: synthesis across processes, places and scales*. Cambridge University Press.

Bosch, J. and Hewlett, J. (1982). "A review of catchment experiments to determine the effect of vegetation changes on water yield and evapotranspiration". In: *Journal of Hydrology* 55.1, pp. 3–23. ISSN: 0022-1694. DOI: `https://doi.org/10.1016/0022-1694(82)90117-2`.

Gupta, H. V., Kling, H., Yilmaz, K. K., and Martinez, G. F. (2009). "Decomposition of the mean squared error and NSE performance criteria: Implications for improving hydrological modelling". In: *Journal of hydrology* 377.1-2, pp. 80–91.

Gustard, A., Bullock, A., and Dixon, J. (1992). *Low flow estimation in the United Kingdom*. Institute of Hydrology.

Gustavsson, H. (2014). *Konsekvenser för avrinning vid metodikbyte mellan HBV och HYPE*.

Häggström, M. (1989). *Anpassning av HBV-modellen till Torneälven*.

He, Y., Bárdossy, A., and Zehe, E. (2011). "A review of regionalisation for continuous streamflow simulation". In: *Hydrology and Earth System Sciences* 15.11, pp. 3539–3553.

Hirsch, R. M. (1979). "An evaluation of some record reconstruction techniques". In: *Water resources research* 15.6, pp. 1781–1790.

Hirsch, R. M. (1982). "A comparison of four streamflow record extension techniques". In: *Water Resources Research* 18.4, pp. 1081–1088.

Hrachowitz, M., Savenije, H., Blöschl, G., McDonnell, J., Sivapalan, M., Pomeroy, J., Arheimer, B., Blume, T., Clark, M., Ehret, U., et al. (2013). "A decade of Predictions in Ungauged Basins (PUB)—a review". In: *Hydrological sciences journal* 58.6, pp. 1198–1255.

Knoben, W. J., Freer, J. E., and Woods, R. A. (2019). "Inherent benchmark or not? Comparing Nash–Sutcliffe and Kling–Gupta efficiency scores". In: *Hydrology and Earth System Sciences* 23.10, pp. 4323–4331.

Länsstyrelsen Västmanlands län (n.d.). *Sävaåns och Hågaåns åtgärdsområden samt Ekolns, Gorrans, Lårstavikens, Skofjärdens och Stora Ullfjärdens närområdens åtgärdsområden-underlag till åtgärdsprogram.*

Leach, J. A. and Laudon, H. (2019). "Headwater lakes and their influence on downstream discharge". In: *Limnology and Oceanography Letters* 4.4, pp. 105–112.

Lindström, G., Johansson, B., Persson, M., Gardelin, M., and Bergström, S. (1997). "Development and test of the distributed HBV-96 hydrological model". In: *Journal of hydrology* 201.1-4, pp. 272–288.

Lindström, G., Pers, C., Rosberg, J., Strömqvist, J., and Arheimer, B. (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.

Marahatta, S., Devkota, L., and Aryal, D. (2021). "Hydrological Modeling: A Better Alternative to Empirical Methods for Monthly Flow Estimation in Ungauged Basins". In: *Journal of Water Resource and Protection* 13.3, pp. 254–270.

McMillan, H. (2020). "Linking hydrologic signatures to hydrologic processes: A review". In: *Hydrological Processes* 34.6, pp. 1393–1409.

McMillan, H. (2021). "A review of hydrologic signatures and their applications". In: *Wiley Interdisciplinary Reviews: Water* 8.1, e1499.

McMillan, H., Westerberg, I., and Branger, F. (2017). "Five guidelines for selecting hydrological signatures". In: *Hydrological Processes* 31.26, pp. 4757–4761.

Olden, J. D. and Poff, N. (2003). "Redundancy and the choice of hydrologic indices for characterizing streamflow regimes". In: *River research and applications* 19.2, pp. 101–121.

Oudin, L., Hervieu, F., Michel, C., Perrin, C., Andréassian, V., Anctil, F., and Loumagne, C. (2005). "Which potential evapotranspiration input for a lumped rainfall–runoff model?: Part 2—Towards a simple and efficient potential evapotranspiration model for rainfall–runoff modelling". In: *Journal of hydrology* 303.1-4, pp. 290–306.

Parajka, J., Viglione, A., Rogger, M., Salinas, J., Sivapalan, M., and Blöschl, G. (2013). "Comparative assessment of predictions in ungauged basins–Part 1: Runoff-hydrograph studies". In: *Hydrology and Earth System Sciences* 17.5, pp. 1783–1795.

Patil, S. and Stieglitz, M. (2011). "Hydrologic similarity among catchments under variable flow conditions". In: *Hydrology and Earth System Sciences* 15.3, pp. 989–997.

Pechlivanidis, I., Crochemore, L., Rosberg, J., and Bosshard, T. (2020). "What are the key drivers controlling the quality of seasonal streamflow forecasts?" In: *Water Resources Research* 56.6, e2019WR026987.

Perrin, C., Oudin, L., Andreassian, V., Rojas-Serna, C., Michel, C., and Mathevet, T. (2007). "Impact of limited streamflow data on the efficiency and the parameters of rainfall—runoff models". In: *Hydrological sciences journal* 52.1, pp. 131–151.

Persson, J., Loreth, T., Johansson, G., and Lönnerholm, Y. (2011). *Fiskinventering i Hågaån 2010*.

Pool, S. and Seibert, J. (2021). "Gauging ungauged catchments–Active learning for the timing of point discharge observations in combination with continuous water level measurements". In: *Journal of Hydrology* 598, p. 126448.

Pool, S., Viviroli, D., and Seibert, J. (2019). "Value of a limited number of discharge observations for improving regionalization: A large-sample study across the United States". In: *Water Resources Research* 55.1, pp. 363–377.

Reynolds, E., Gustafsson, D., and Clemenzi, I. (2021). *Spring Flood Predictions with HBV and HYPE*.

Seibert, J. (1999). "Regionalisation of parameters for a conceptual rainfall-runoff model". In: *Agricultural and forest meteorology* 98, pp. 279–293.

Seibert, J. (2000). "Multi-criteria calibration of a conceptual runoff model using a genetic algorithm". In: *Hydrology and Earth System Sciences* 4.2, pp. 215–224.

Seibert, J. and Bergström, S. (2021). "A retrospective on hydrological modelling based on half a century with the HBV model". In: *Hydrology and Earth System Sciences Discussions*, pp. 1–28.

Seibert, J. and Beven, K. J. (2009). "Gauging the ungauged basin: how many discharge measurements are needed?" In: *Hydrology and Earth System Sciences* 13.6, pp. 883–892.

Seibert, J. and Vis, M. J. (2016). "How informative are stream level observations in different geographic regions?" In: *Hydrological Processes* 30.14, pp. 2498–2508.

Sjökvist, E. and Abdoush, D. (2019). *Sommaren 2018-en glimt av framtiden?*

SMHI (2018). *Kalibrering och utvärdering av S-HYPE*. URL: `https://www.smhi.se/data/hydrologi/vattenwebb/om-data-i-vattenwebb/kalibrering-och-utvardering-av-s-hype-1.26735`.

SMHI (2020). *S-HYPE: HYPE-modell för hela Sverige*. URL: `https://www.smhi.se/forskning/forskningsenheter/hydrologisk-forskning/s-hype-hype-modell-for-hela-sverige-1.560`.

SMHI (2022a). *Modelldata per område*. URL: `https://vattenwebb.smhi.se/modelarea/`.

SMHI (2022b). *SMHI mätningar*. URL: `https://vattenwebb.smhi.se/station/`.

SMHI (2022c). *Utvärdera modellresultat för sötvatten*. URL: `https://vattenwebb.smhi.se/modeldiff/`.

SMHI (n.d.). *PTHBV Klimatdatabas för hydrologiska beräkningar*. URL: `https://www.smhi.se/polopoly_fs/1.166963!/PTHBV_produktblad%5C%20DM19775.pdf`.

Sorooshian, S., Gupta, V. K., and Fulton, J. L. (1983). "Evaluation of maximum likelihood parameter estimation techniques for conceptual rainfall-runoff models: Influence of calibration data variability and length on model credibility". In: *Water Resources Research* 19.1, pp. 251–259.

Spearman, C. (1904). "The proof and measurement of association between two things". In: *The American Journal of Psychology* 15.1, pp. 72–101.

Tobler, W. R. (1970). "A computer movie simulating urban growth in the Detroit region". In: *Economic geography* 46.sup1, pp. 234–240.

Van Kempen, G., Van Der Wiel, K., and Melsen, L. A. (2021). "The impact of hydrological model structure on the simulation of extreme runoff events". In: *Natural Hazards and Earth System Sciences* 21.3, pp. 961–976.

Personal Communication

Fölster, J. (2022). *Department of Aquatic Sciences and Assessment, Swedish University of Agricultural Sciences.*

Östlund, M. (2022). *Department of Aquatic Sciences and Assessment, Swedish University of Agricultural Sciences.*

Sandehed, G. (2022). *SMHI.*

Seibert, J. (2022). *Department of Geography, University of Zurich.*

Vis, M. (2022). *Department of Geography, University of Zurich.*

8 Appendix

8.1 Differances in weather between "calibration" and evaluation period

Table 7. Some differances between climate/weather in the different catchments for the "calibration period" (1981-1990, i.e. the period used to calculate mean and standard deviation for the MOVE method, as well as the period used for calibrating UB), and the evaluation period (1991-2000). Only four of the catchments are in the table as meteorological data was only available for those catchments.

	Q [mm/year]	Precipitation [mm/year]	Q/P	ET
Håga calibration	333	656	0.51	323
Håga evaluation	304	599	0.51	295
Stabby calibration	268	638	0.42	370
Stabby evaluation	189	584	0.32	395
Säva calibration	291	661	0.44	371
Säva evaluation	233	609	0.38	376
Örsunda calibration	302	667	0.45	366
Örsunda evaluation	239	622	0.38	383

8.2 Point observations of streamflow

The following figures show the actual point observations from the hydrological year 2018 in relation to the water-level time-series, as well as the "synthectic" streamflow observations from the evaluation period in relation to the full streamflow record. The actual values of the observations (in mm/day) is 0.015, 0.72 and 1.92 for the modern period, and 0.028, 0.87 ans 2.31 for the old, "synthetic", observations.

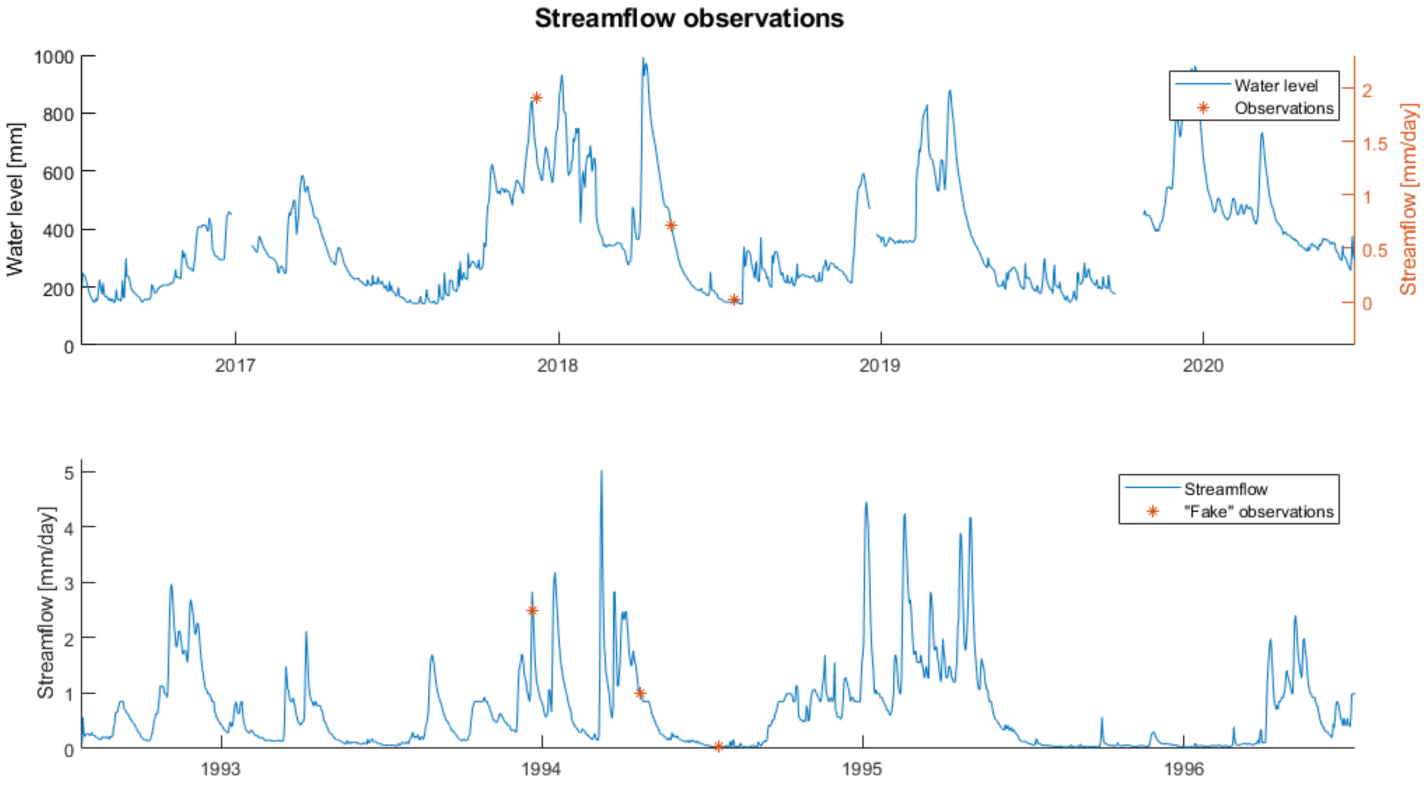

Figure 13. *Point observations of streamflow and their corresponding time-series*

8.3 Ranges for HBV-parameters

Table 8. The values and descriptions are taken from the article by Seibert et al. (2016), except for the value for K2 which was discussed with Seibert in a meeting (Seibert 2022).

Parameter	Description	Min	Max	Unit
Snow routine				
TT	Threshold temperature	-2	2.5	$°C$
CFMAX	Degree-day factor	0.5	10	$mm \cdot °C^{-1} \cdot d^{-1}$
SFCF	Snowfall correction factor	0.5	1.2	-
CFR	Refreezing coefficient	0	0.1	-
CWH	Water holding capacity	0	0.2	-
Soil Routine				
FC	Maximum storage in soil box	100	550	mm
LP	Threshold for reduction of evaporation	0.3	1	-
BETA	Shape coefficient	1	5	-
Response routine				
PERC	Maximal flow from upper to lower box	0	4	$mm \cdot d^{-1}$
ALPHA	Non-linearity coefficient	0	70	-
K1	Recession coefficient (upper box)	0.01	0.2	d^{-1}
K2	Recession coefficient (lower box)	0.001	0.1	d^{-1}
Routing routine				
MAXBAS	Routing, length of weighting function	1	5	d

8.4 Rating curve

Rating curves are usually constructed using the following equation:

$$Q = C(H - H_0)^n \tag{10}$$

Where Q is the discharge, H is the water stage, H_0 is the water stage that corresponds to no flow, while C and n are empirical constants. H_0, C and n thus have to be found to construct a rating equation. Through visually matching a curve to the three available streamflow point observations, combined with knowledge of flow magnitude and common water-levels, the following equation was found:

$$Q = 9(H - 0.14)^{1.7} \tag{11}$$

Which can be visualised as a rating curve as seen in Figure 14

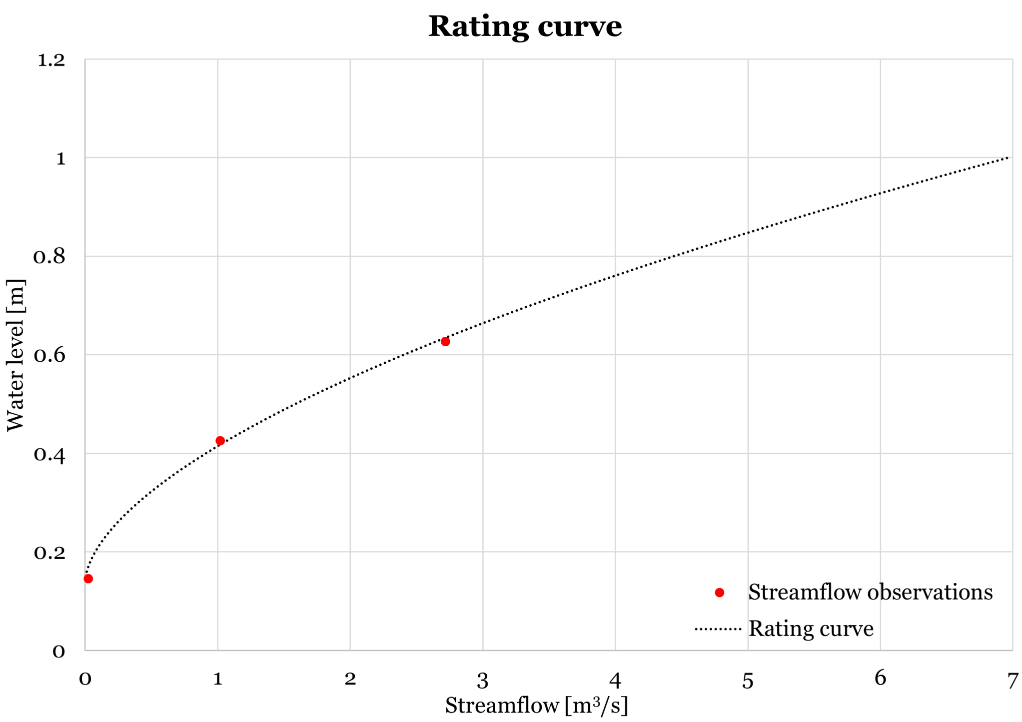

Figure 14. *The rating curve for Håga catchment created for this project.*

8.5 KGE-values

Table 9. The KGE-values for the different daily streamflow estimations, as well as the seasonal estimations. The performance measured by Nash–Sutcliffe efficiency (NSE) is also supplied, to be used as a reference when comparing results to other studies. For the HBV-method the number refers to the KGE/NSE-value of the ensemble mean.

METHOD	Daily NSE	Daily KGE	Seasonal KGE			
			Winter	Spring	Summer	Fall
Index						
DAR						
Säva	0.76	0.71	0.59	0.79	0.77	0.76
Sävja	0.49	0.37	0.27	0.49	0.14	0.35
Stabby	0.52	0.58	0.49	0.65	0.31	0.59
Vattholma	0.30	0.36	0.16	0.50	0.01	0.24
Örsunda	0.57	0.71	0.70	0.71	0.51	0.60
MOVE						
Säva	0.78	0.85	0.76	0.80	0.80	0.83
Sävja	0.67	0.61	0.47	0.76	0.31	0.57
Stabby	0.72	0.76	0.60	0.85	0.66	0.87
Vattholma	0.33	0.46	0.25	0.61	0.03	0.30
Örsunda	0.37	0.63	0.71	0.51	0.38	0.34
Model						
S-HYPE	0.71	0.72	0.54	0.78	0.62	0.90
HBV						
UB	0.73	0.76	0.69	0.75	0.63	0.83
LB	0.40	0.40	0.17	0.54	0.36	0.35
waterlevel	0.61	0.62	0.53	0.57	0.14	0.62
Waterlevel_MSE	0.60	0.62	0.55	0.58	0.14	0.63
Point_Obs ("modern")	0.56	0.52	0.42	0.54	0.12	0.41
Point_Obs ("old")	0.68	0.62	0.52	0.69	0.50	0.61
Rating curve	0.69	0.72	0.70	0.76	0.14	0.58
Donor catchment	0.67	0.66	0.60	0.77	0.38	0.58

8.6 Distribution of KGE for UB and LB

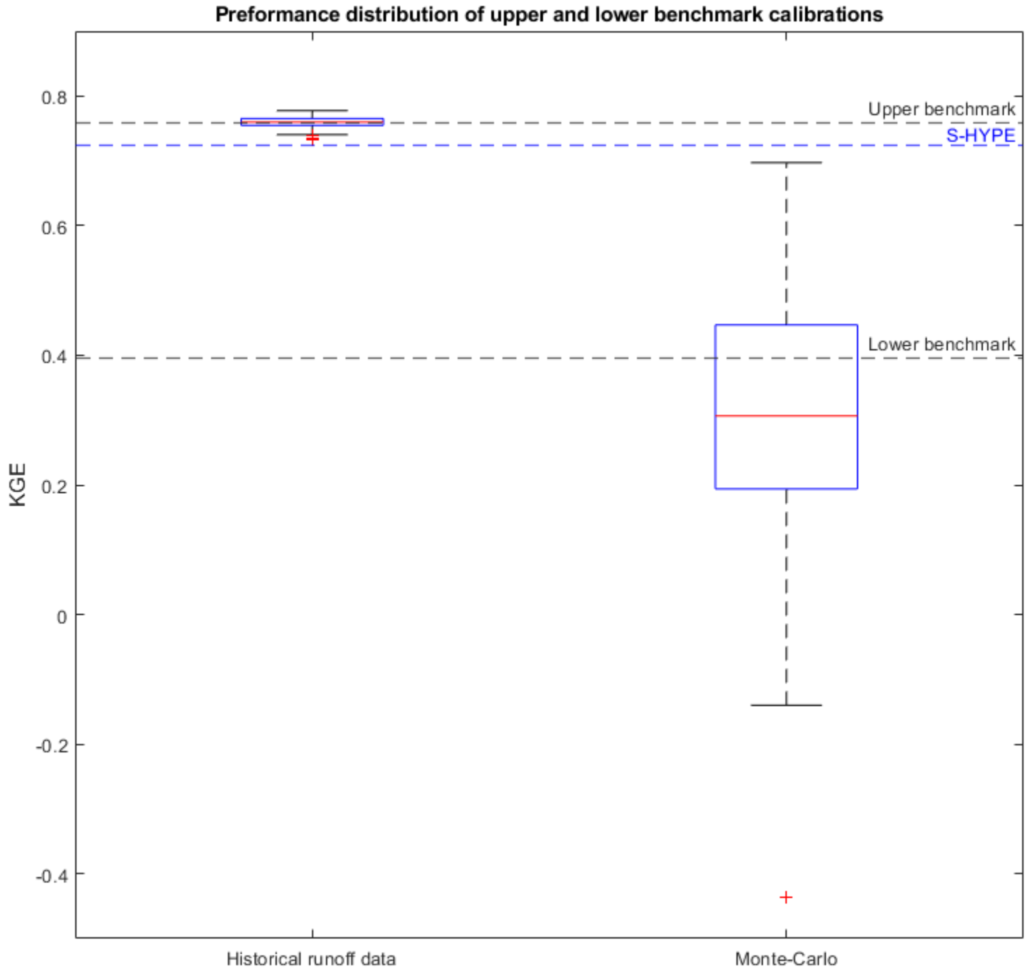

Figure 15. *performance distribution of UB and LB*

8.7 Hydrological signatures for calibration with old and new data

Table 10. Hydrological signatures for the HBV calibration-method using old and modern discharge point observations.

	Spearman	$Q_{diff}[mm/yr]$	FI	BFI
Håga	1	0.00	0.12	0.64
Point_Obs ("modern")	0.86	-86	0.06	0.75
Point_Obs ("old")	0.91	-63	0.10	0.65

Table 11. Hydrological signatures related to high- and low-flows for the HBV calibration-method using old and modern discharge point observations.

	MLQ $[mm/day]$	MHQ $[mm/day]$	Bias Q95	Bias Q90	Bias Q10	Bias Q5	Freq. lowflow $[d/yr]$	Freq. highflow $[d/yr]$
Håga	0.15	4.14	0	0	0	0	65	12
Point_Obs ("modern")	0.08	2.12	0.45	0.06	-0.45	-0.51	84	1
Point_Obs ("old")	0.11	3.24	0.64	0.42	-0.36	-0.39	48	3