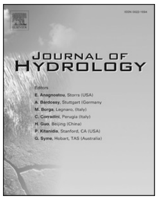
Contents lists available at ScienceDirect

Journal of Hydrology

journal homepage: www.elsevier.com/locate/jhydrol



Research papers

Identifying major drivers of daily streamflow from large-scale atmospheric circulation with machine learning

Jenny Sjåstad Hagen[a,b,*], Etienne Leblois[c], Deborah Lawrence[d], Dimitri Solomatine[e], Asgeir Sorteberg[a,b]

[a] University of Bergen, Geophysical Institute, Allegaten 70, 5007 Bergen, Norway
[b] University of Bergen, Bjerknes Centre for Climate Research, Jahnebakken 5, 5007 Bergen, Norway
[c] French National Institute for Agriculture, Food, and Environment (INRAE), Riverly-Lyon Research Unit, 5 rue de la Doua, 69625 Villeurbanne, France
[d] Norwegian Water Resources and Energy Directorate (NVE), Middelthuns gate 29, 0368 Oslo, Norway
[e] IHE Delft Institute for Water Education, Westvest 7, 2611 AX Delft, Netherlands

ARTICLE INFO

Keywords:
Direct downscaling
Discharge
Automated feature extraction
ERA5
ERA-Interim

ABSTRACT

Previous studies linking large-scale atmospheric circulation and river flow with traditional machine learning techniques have predominantly explored monthly, seasonal or annual streamflow modelling for applications in direct downscaling or hydrological climate-impact studies. This paper identifies major drivers of daily streamflow from large-scale atmospheric circulation using two reanalysis datasets for six catchments in Norway representing various Köppen-Geiger climate types and flood-generating processes. A nested loop of roughly pruned random forests is used for feature extraction, demonstrating the potential for automated retrieval of physically consistent and interpretable input variables. Random forest (RF), support vector machine (SVM) for regression and multilayer perceptron (MLP) neural networks are compared to multiple-linear regression to assess the role of model complexity in utilizing the identified major drivers to reconstruct streamflow. The machine learning models were trained on 31 years of aggregated atmospheric data with distinct moving windows for each catchment, reflecting catchment-specific forcing-response relationships between the atmosphere and the rivers. The results show that accuracy improves to some extent with model complexity. In all but the smallest, rainfall-driven catchment, the most complex model, MLP, gives a Nash-Sutcliffe Efficiency (NSE) ranging from 0.71 to 0.81 on testing data spanning five years. The poorer performance by all models in the smallest catchment is discussed in relation to catchment characteristics, sub-grid topography and local variability. The intra-model differences are also viewed in relation to the consistency between the automatically retrieved feature selections from the two reanalysis datasets. This study provides a benchmark for future development of deep learning models for direct downscaling from large-scale atmospheric variables to daily streamflow in Norway.

1. Introduction

The plethora of available large-scale atmospheric reanalysis products, climate model outputs and weather (re-)forecasts is growing; in combination with quality-controlled streamflow records, this growth unlocks unprecedented opportunities for streamflow modelling with machine learning on various spatiotemporal scales. The use of machine learning in hydrology has evolved over the last decades in association with the field of Hydroinformatics (Abbott, 1991; Vojinovic and Abbott, 2017). A historical overview of the use of neural networks in hydrology is given by Abrahart et al. (2012). A review of machine learning applications for streamflow modelling is given by Yaseen et al. (2015), while Mosavi et al. (2018) provide a review of machine learning applications specifically for flood prediction. In recent years, a growing interest in deep learning is reflected in the literature (Kratzert et al., 2018; Kratzertet al., 2019; Xiang et al., 2020), and attempts at integrating physical laws of hydrology in deep learning models have been made (Jiang et al., 2020). Meanwhile, traditional machine learning techniques, such as random forest, support vector machine and shallow neural networks, have demonstrated successful applications in the field of hydrology.

Nevertheless, machine learning applications for streamflow

[*] Corresponding author at: Geophysical Institute, University of Bergen, Jahnebakken 5, 5007 Bergen, Norway.
E-mail address: jenny.hagen@uib.no (J.S. Hagen).

https://doi.org/10.1016/j.jhydrol.2021.126086
Received 16 November 2020; Received in revised form 28 January 2021; Accepted 13 February 2021
Available online 23 February 2021

Table 1
Overview of studies predicting streamflow from large-scale atmospheric circulation (see Table A1 in Appendix A for list of symbols). (See below-mentioned references for further information.)

Article	Objective	Country	Machine learning technique(s)	Input selection procedure	Input (from Global Circulation Model)				Output
					Variables				
					Spatial resolution (º)	*Temporal resolution (days)*	*Pressure level (hPa)*	**Catchment characteristics*	
Rasouli et al. (2012)	Short-term streamflow modelling	USA	Supervised (SVM, BNN)	Stepwise predictor selection	Climate indices; weather forecasts				Daily streamflow
					t2m, SLP, Precip, Tcw, RH, Heat, U, V (wind phase and amplitude)				
					2.0 x 2.0	½ (12-hourly)	500; 700; 1000	No	
Ghosh and Mujumdar (2008)	Climate-impact streamflow modelling	India	Supervised (SVM, RVM)	PCA, fuzzy clustering	Reanalysis data; climate model outputs				Monthly (monsoon) streamflow
					Z, SLP, t2m, SH				
					2.5 x 2.5 (5.5 x 6.25)	30 (monthly)	500; 1000	No	
Cannon and Whitfield (2002)	Short-term streamflow modelling	Canada	Supervised (MLP)	Stepwise predictor selection	Reanalysis data				Five-day-average streamflow
					Z, SLP, SH				
					2.5 x 2.5 (3.75 x 3.75)	5 (five-daily)	500; 850; 1000	No	
Tisseuil *et al.* (2010)	Climate-impact streamflow modelling	France	Supervised (MLP)	Expert knowledge/ prior assumptions	Reanalysis data; climate model outputs				Daily streamflow and fortnightly flow statistics
					T, Precip, Z, RH, SH, SST, SLP, LHeat, SHeat, SW, LW, Cloud				
					2.5 x 2.5	1; 14 (daily)	500; 850; 1000	No	
Moradi *et al.* (2020)	Long-term streamflow modelling	Iran	Supervised (MLP, ANFIS)	Correlation analysis, PCA	Climate indices; reanalysis data				Annual streamflow
					SLP, T, Z, RH, Ω				
					2.5 x 2.5	30 (monthly)	500; 600; 700; 850; 1000	No	
Okkan and Inan (2015)	Climate-impact streamflow modelling	Turkey	Supervised (SVM, RVM)	PCA	Reanalysis data; climate model outputs				Monthly streamflow
					T, Z, RH, Precip, SLP				
					2.5 x 2.5	30 (monthly)	200; 500; 850; 1000	No	
Sachindra *et al.* (2013)	Climate-impact streamflow modelling	Australia	Supervised (SVM)	Correlation analysis	Reanalysis data (used as proxy for climate model outputs)				Monthly streamflow
					RH, SH, T, Z				
					2.5 x 2.5	30 (monthly)	500; 700; 850; 1000	Yes (soil moisture)	
Ren *et al.* (2018)	Climate-impact streamflow modelling	China	Supervised (BNN, SVM)	Expert knowledge/ prior assumptions	Reanalysis data; climate model outputs				Monthly streamflow
					t2m, SLP, Z, RH, SH, Precip, U, V				
					1.25 x 1.25	30 (monthly)	500; 850; 1000	No	
Nilsson *et al.* (2008)	Seasonal streamflow modelling	Norway, Sweden	Supervised (MLP)	Canonical correlation analysis	Weather forecasts				Bimonthly streamflow
					U, moisture variates (not specified)				
					2.8 x 2.8	90 (3-monthly)	850	No	
Liao *et al.* (2019)	Short-term streamflow modelling	China	Supervised (MLP, SVM, GBRT)	Maximal information coefficient	Reanalysis data				Daily streamflow
					Precip, td2m, LW, t2m, Tcv, Tcw				
					0.75 x 0.75	¼ (6-hourly)	1000	Yes (soil moisture/ temperature; snow layer)	
Sahoo and Sen (2017)	Climate-impact streamflow modelling	USA	Supervised (SVM)	PCA, fuzzy clustering	Reanalysis data; climate model outputs				Maximum seasonal streamflow
					U, V, RH, Tcw, T				
					2.5 x 2.5 (3.75 x 3.75)	30 (monthly)	1000	No	
Das and Nanduri (2018)	Climate-impact streamflow modelling	India	Supervised (SVM, RVM)	PCA, fuzzy clustering	Reanalysis data; climate model outputs				Monthly (monsoon) streamflow
					SLP, SH, Z, t2m				
					2.5 x 2.5	30 (monthly)	500; 1000	No	
Chu et al. (2020)	Long-term streamflow modelling	USA	Supervised (DBN)	Least absolute shrinkage and selector operator	Climate indices; reanalysis data				Monthly streamflow
					SST				
					1.0 x 1.0	30 (monthly)	1000	No	
(Huang *et al.* (2020)	Climate-impact streamflow modelling	Norway	Supervised (RVM)	PCA, correlation analysis	Reanalysis data				Monthly streamflow
					t2m, T, Z, SLP, RH, SH				
					2.5 x 2.5	30 (monthly)	200; 500; 700; 850; 1000	Yes (snow water equivalent)	

**Catchment characteristics refer to any of the following: local observations of surface variables, catchment-specific parameters or characteristics affecting runoff, such as land cover, soil, slope etc.*

modelling predominantly take the form of autoregressive models, in which past streamflow observations represent the primary input variables (Adnan et al., 2019; Thapa et al., 2020; Tongal and Booij, 2018) – sometimes in combination with local meteorological observations, climate indices and weather forecasts (see for instance Rasouli et al. (2012)). Few studies have utilized machine learning for direct translation from large-scale atmospheric circulation to streamflow without also employing past streamflow observations as predictive variables.

It has long been known that there are important linkages between large-scale atmospheric circulation and river flow (Kingston et al., 2006); Table 1 provides an overview of relevant studies in which streamflow has been inferred from gridded large-scale atmospheric variables. An early attempt at modelling five-day-average streamflow from atmospheric reanalysis data with ensemble neural networks is presented by Cannon and Whitfield (2002). Feature selection was carried out with stepwise predictor selection, resulting in the use of 2.5° x 2.5° (~250 km × 250 km) gridded cells of geopotential heights, mean sea level pressure and specific humidity at 500 hPa, 850 hPa and 1000 hPa. The study concluded that neural networks are more suitable than stepwise linear regression for nonlinear systems with complex interactions between inputs and outputs. Following this, a number of studies used machine learning to reconstruct monthly streamflow and project hydrological climate-impacts.

Ghosh and Mujumdar (2008) trained support vector machine (SVM) and relevant vector machine (RVM) models on atmospheric reanalysis data to reconstruct monthly streamflow in the monsoon season. The trained models were subsequently used to project hydrological climate-impacts with climate model outputs interpolated to the grid points of the reanalysis data. Feature selection was performed on a prior selection of large-scale atmospheric variables with known physical connections to the monsoon-flooding regime. Principal component analysis (PCA) was carried out on 2.0° x 2.0° (~200 km × 200 km) gridded cells of standardized mean sea level pressure, air temperature, specific humidity and 500 hPa geopotential to transform the correlated predictors into an *n*-dimensional set of uncorrelated vectors, whereby most of the variability of the original dataset is contained within the first few dimensions. Fuzzy clustering of the principal components was performed using three classes. The resulting 12 inputs comprised 10 principal components and two memberships (as the sum of three memberships is one). Similar approaches to feature extraction with PCA and fuzzy clustering for hydrological climate-impact modelling with SVM have been followed by Okkan and Inan (2015), Sahoo and Sen (2017) and Das and Nanduri (2018).

In the case of Norway, monthly streamflow has been directly downscaled from moving windows of atmospheric reanalysis data by Huang et al. (2020) using RVM. Precipitation was excluded as an input variable in order to make use of variables that are more accurately simulated by global circulation models (temperature, pressure and humidity). The best performance was obtained in snowmelt-driven catchments in inland regions of Norway. The poor performance in small, rainfall-driven catcments was associated with the relative difference in inter-annual versus intra-annual variability in monthly streamflow.

As is evident from Table 1, most attempts at direct downscaling or climate-impact modelling with machine learning have targeted monthly, seasonal or annual streamflow, often using variations of correlation analysis for feature selection. Ren et al. (2018) justified selected features with expert knowledge and prior assumptions to assess climate-impact on *monthly* streamflow in China. Nilsson et al. (2008) used canonical correlation analysis to extract features for *seasonal* streamflow modelling with MLP in Scandinavia. Moradi et al. (2020) extracted relevant features using correlation analysis and PCA for *annual* streamflow modelling with MLP to a reservoir in the Middle East. The majority of the previous studies on monthly, seasonal or annual streamflow modelling have used monthly aggregations of atmospheric inputs.

On finer temporal scales, few attempts can be found in the literature. Tisseuil et al. (2010) modelled daily mean flow and fortnightly flow statistics from 2.5° x 2.5° (~250 km × 250 km) gridded atmospheric reanalysis data and projected hydrological climate-impact under two scenarios with neural networks. Almost a decade later, Liao et al. (2019) employed only surface variables from 0.75° x 0.75° (~75 km × 75 km) gridded reanalysis data to model daily streamflow and assess the hydrological impact of climate change using climate model outputs. Overall, despite increasing spatiotemporal resolution of available reanalysis data, attempts at reconstructing daily streamflow from large-scale atmospheric circulation with machine learning remain sparse.

Reichstein et al. (2019) identified five main challenges for successful application of deep learning in geosciences: i) *interpretability* (self-explanatory and transparent model structures that allow for causal discovery from observational data), ii) *physical consistency* (integration of domain knowledge), iii) *complex and uncertain data* (methods to cope with complex statistics and sources of uncertainty), iv) *limited labels* (methods to exploit data with unsupervised or semi-supervised learning) and v) *computational demand* (technical challenges of Big Data computing). However, recognizing the lack of non-autoregressive machine learning models for direct downscaling to daily streamflow in the literature, and embracing the fact that a model should be as simple as possible, but nevertheless reliable, this paper focuses on the two first points highlighted by Reichstein et al. (2019), namely *interpretability* and *physical consistency*, within the context of traditional machine learning techniques. Hence, a thorough investigation of deep learning is reserved for future studies.

While other studies have focused primarily on monthly, seasonal or annual streamflow modelling, this study focuses on the identification of major drivers of *daily* streamflow among relevant atmospheric and surface variables in relation to distinct flood-generating processes. The objective of this paper is to identify major drivers and investigate the forcing-response relationship between large-scale atmospheric circulation and daily streamflow with traditional machine learning techniques of increasing model complexity. Specifically, this paper answers the following research question: What are major drivers of daily streamflow in six selected Norwegian catchments with different climate types, flood-generating processes and catchment characteristics, and how does the direct translation from atmospheric forcing to streamflow response relate to model complexity? In other words, this paper provides a sampled screening of the potential of direct downscaling from large-scale atmospheric variables to streamflow at a daily temporal scale in Norway. The aim is to present a methodology for identification of major drivers of daily streamflow and assess the accuracy of reconstructed streamflow in relation to model complexity. Major drivers are identified using an automatic feature selection procedure that returns atmospheric variables deemed relevant at informative pressure levels and aggregations in the atmospheric column above the relevant catchment. Three popular traditional machine learning techniques with differing structural complexities are explored: the ensemble-based, piecewise linear random forest (RF); the nonlinear, kernel-based support vector machine (SVM) for regression; and multilayer perceptron (MLP) neural network. Multiple-linear regression is used as a baseline model for the comparisons.

This paper is structured as follows. The remainder of Section 1 introduces the case study (1.1). Section 2 presents materials and methods in subsections of data (2.1), feature selection (2.2), model development (2.3) and evaluation (2.4). Results are shown in Section 3. A discussion follows in Section 4, with three subsections focusing on model performance (4.1), interpretability and physical consistency (4.2) and potential and limitations (4.3) respectively. Lastly, Section 5 provides concluding remarks and gives directions for future research.

1.1. Case study

With 1660 hydropower plants accounting for more than 95% of

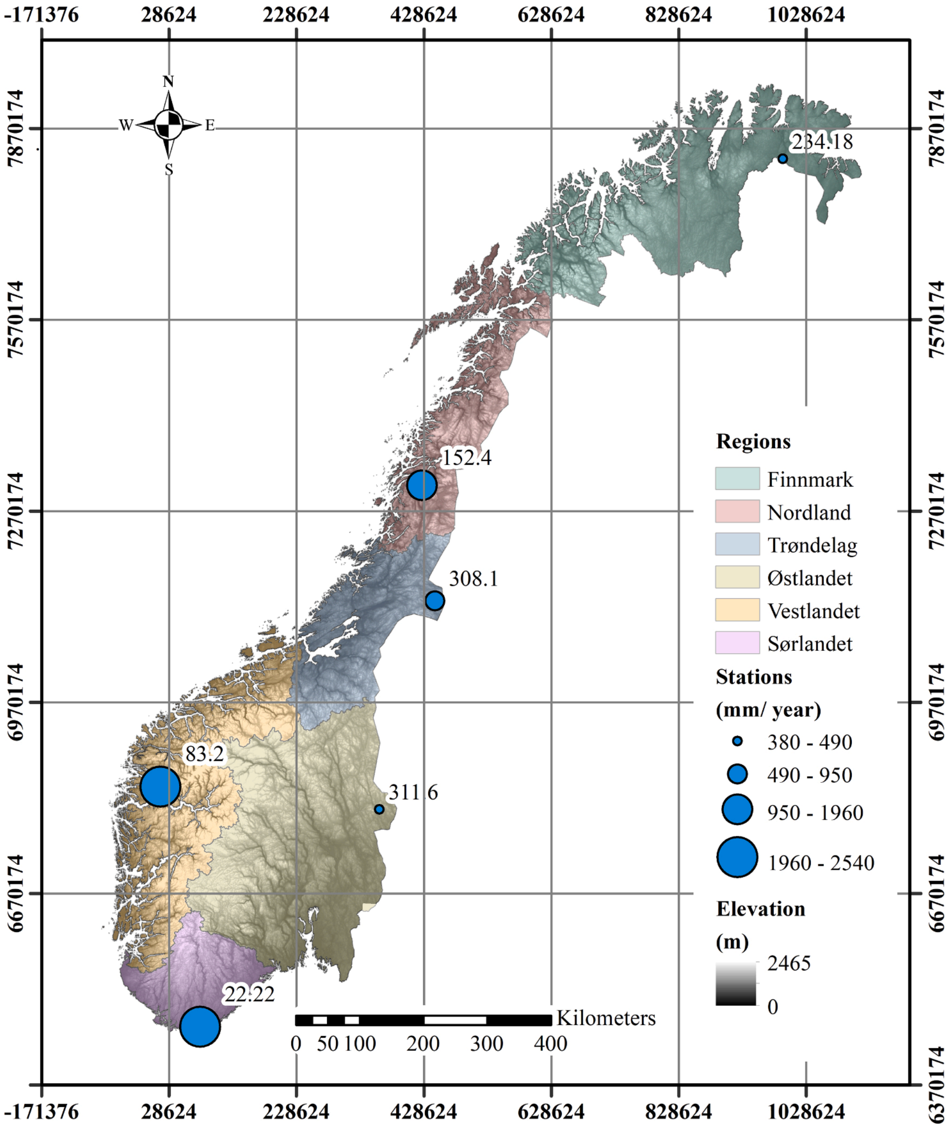


Fig. 1. Overview of station locations and regions. The stations are scaled by mean annual flow divided by catchment area. Coordinate system: WGS84. Projection: UTM 33 N. Digital elevation data retrieved from: http://www.viewfinderpanoramas.org/Coverage map viewfinderpanoramas_org1.htm.

installed capacity, there are numerous regulated rivers in Norway (Norwegian Ministry of Petroleum and Energy, 2015). However, when linking large-scale atmospheric circulation and river flow with machine learning, it is pivotal that unregulated rivers are used, so that signals from atmospheric forcing are less distorted by human activity. The Norwegian Hydrological Reference Dataset for Climate Impact Studies (Fleig et al., 2013) consists of daily records from 188 active and unregulated streamflow gauges deemed suitable for climate impact studies. From these 188 stations, 40 stations distributed over six regions of Norway were extracted as potential candidates for this study based on a filtering procedure considering record lengths, record gaps, flow duration curves and mean annual flows from 1950 and onwards. The resulting dataset contained four stations in Finnmark; five stations in Nordland; eight stations in Trøndelag; ten stations in Vestlandet; ten stations in Østlandet; and three stations in Sørlandet. Six representative stations (one for each region: Finnmark, Nordland, Trøndelag, Østlandet, Vestlandet, Sørlandet) were selected based on the correlation matrices for each region, on the grounds of having the overall highest correlation with other stations in the region. Fig. 1 shows the geographical location of the six selected hydrological stations at the catchment outlets and the corresponding annual mean flow normalized by catchment area. Important hydro-climatic characteristics for each catchment are given in Table 2.

Station 234.18 is located in the Tana River Basin and is the northernmost catchment considered in this study. This catchment has clearly the most pronounced snowmelt-driven flood regime. Station 152.4 measures the outflow from Lake Fustvatn, which discharges further downstream into the fjord Vefsnfjorden through the river Fusta. Both rainfall and snowmelt contribute to flood generation in this catchment, although rainfall has a tendency to dominate over snowmelt due to the coastal climate. Station 308.1 measures the flow in Lake Lenglingen, which is part of a network of lakes in mid-inland Norway. The main flood-generating process is snowmelt in this catchment. Station 311.6 measures the river flow at Nybergsund, which is part of a larger river system that ultimately discharges into Lake Vånern in Sweden. The main flood-generating process is snowmelt – albeit marginally, as mixed processes also play an important role. The station 83.2 is located in the coastal west, and the main flood-generating process is rainfall. The most distinct rainfall-driven flood regime is found in the southernmost catchment: Station 22.22 measures river flow in Søgne, which connects to the fjord Høllefjorden and eventually drains to the North Sea.

The selected catchments differ with respect to size, climate types and flood-generating processes. The catchments in Vestlandet and Sørlandet are located in coastal regions with a temperate oceanic climate.

Table 2
Overview of stations and catchment characteristics, with flood-generating processes as identified in past studies.

Characteristics	Station/ Region					
	234.18/ Finnmark	*152.4/ Nordland*	*308.1/ Trøndelag*	*311.6/ Østlandet*	*83.2/ Vestlandet*	*22.22/ Sørlandet*
Coordinates	70.03484°N,27.99344°E	65.90595°N,13.36730°E	64.28160°N,13.88369°E	61.32781°N,12.36317°E	61.36779°N,5.91008°E	58.08808°N,7.82599°E
Catchment area (km^2)	14 161	526	450	4 425	508	204
Mean flow 1950–2010 (m^3/s)	170.8	32.6	13.5	69.1	41.0	15.0
Mean annual precipitation(mm)	494	2344	797	735	2304	1963
Flood generating process (%)(Vormoor et al., 2015)	Snowmelt(-)	Rainfall (66%), snowmelt (20%)	Snowmelt (89%), mixed (8%)	Snowmelt (67%), mixed (21%)	Rainfall (62%), snowmelt (26%)	Rainfall(-)
Average fraction rainfall (Engeland et al., 2016)	23%	67%	25%	30%	76%	95%
Köppen-Geiger climate classification (class)(Kottek et al., 2006)	Subarctic climate – no dry season, cold summer (Dfc)	Subarctic climate – no dry season, cold summer (Dfc)	Subarctic climate – no dry season, cold summer (Dfc)	Subarctic climate – no dry season, cold summer(Dfc)	Temperate oceanic climate – no dry season, warm summer(Cfb)	Temperate oceanic climate – no dry season, warm summer (Cfb)
Other remarks	–	Break in flow duration curve (low flows)	Sparse high flows	–	Water has been led out from a 1 km^2 sub-catchment since 1960	–

However, given the tilt of the mountain range near the western coast of Norway, winds carrying moisture eastward and northward have differing effects on the two catchments. All other catchments are located in areas with a subarctic climate, where both snowmelt and rainfall contribute to generation of high flows to various degrees. Similar characteristics are found for the selected catchments in Finnmark and Trøndelag. The catchment in Østlandet exhibits a distinctly different hydrological behavior, in which mixed processes play a larger role, although this station generally exhibits characteristics similar to the stations in Finnmark and Trøndelag. Likewise, the catchment in Nordland differs from the other catchments by being a high-latitude predominantly rainfall-driven catchment, with some contribution from snowmelt; in terms of hydrological behavior, this catchment resembles the catchments in Vestlandet and Sørlandet. In terms of catchment size, the smallest catchment (204 km^2) is found in Sørlandet, and the largest catchment (14 161 km^2) is found in Finnmark (see Table 2).

2. Materials and methods

2.1. *Data*

2.1.1. *Hydrological data*

The flow duration curves of the six selected stations are shown in Fig. 2. Mann-Kendall tests (Hussain and Mahmud, 2019) of maximum annual flows and average annual flows show no statistically significant monotonically increasing trends between 1979 and 2018 (see Table 3). Descriptive statistics are summarized Table 4.

2.1.2. *Atmospheric data*

Atmospheric data was retrieved from two reanalysis datasets of differing spatial resolution: a fine-resolution reanalysis dataset, ECMWF-ERA5 (Hersbach et al., 2020), and a coarser-resolution reanalysis dataset, ECMWF ERA Interim (hereby abbreviated ERAI) (Dee et al., 2011). The column above each hydrological station was downloaded for a standard selection of variables at 20 pressure levels between 1000 hPa and 50 hPa. The atmospheric reanalysis data were aggregated to daily values by summing or averaging extensive and intensive variates, respectively. Table 5 provides an overview of the atmospheric reanalysis data and variates used for feature selection (see Table A1 in Appendix A for list of symbols). Table 6 shows the geographical coordinates of the atmospheric columns.

2.2. *Feature selection*

The hydrological and atmospheric data spans the period January 1, 1979, to December 31, 2017. To ensure a testing set of five consecutive years and also reserve the maximum flows at each hydrological station for the training set, the data was split as shown in Fig. 3. Two years (1979–1981) were set aside for cross-validation to prevent overfitting. Five years (2009–2014) were reserved for testing. The remaining 31 years were used for training.

The simplest machine learning model considered in this study, random forest (RF), was first used to select features as identical input variable selections to all machine learning models as well as the baseline model. Feature selection was carried out on the training data with two loops of random forests based on the following assumptions:

1. Streamflow is, albeit to various degrees, directly forced by large-scale atmospheric circulation as represented in the gridded pressure columns above the catchments;

2. Aggregations and/ or lags of atmospheric variables can only influence present streamflow if aggregated backwards in time;

3. Catchment response times and hydrological regimes vary from station to station even within a near-homogenous region. Therefore, each station requires a unique set of aggregated atmospheric inputs to establish a representation of catchment characteristics in the forcing-response relationship.

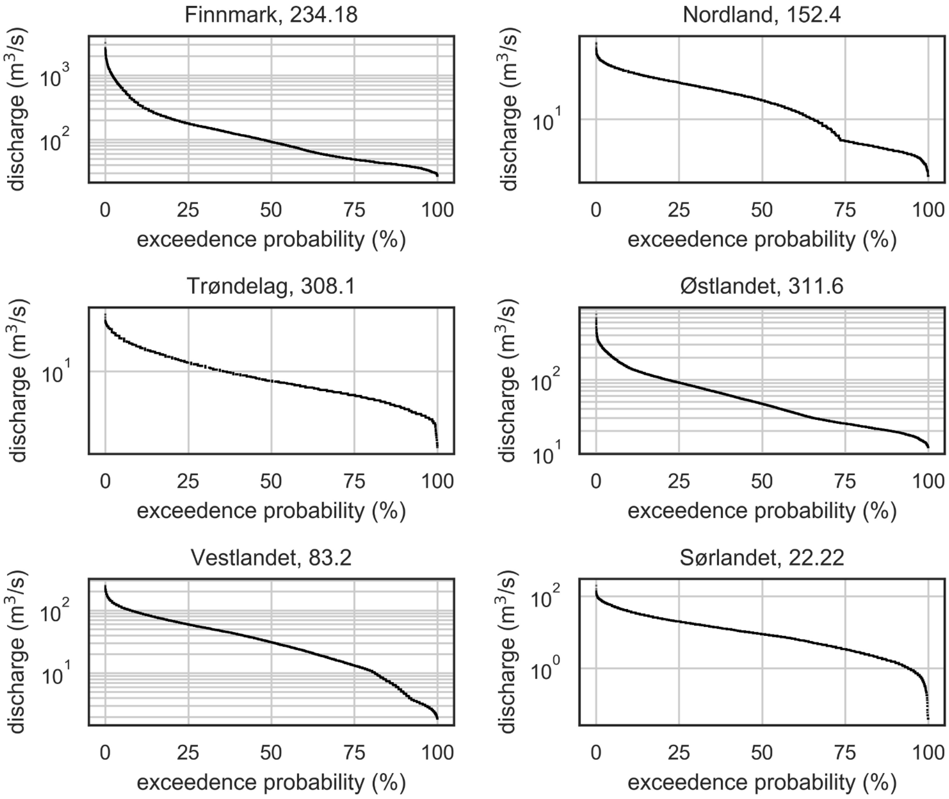


Fig. 2. Flow duration curves with logarithmic scales.

Table 3
Mann-Kendall trend analysis (only shown for maximum annual flow) at 0.05 significance level. z = normalized test statistic, s = Mann Kendall's score, p = p-value of the significance test.

	Slope (m^3/yr)	z	s	p	Trend
234.18/ Finnmark	−5.46	−0.60	−51	0.55	No trend
152.4/ Nordland	0.69	1.26	105	0.21	No trend
308.1/ Trøndelag	−0.09	−0.31	−27	0.75	No trend
311.6/ Østlandet	−0.85	−0.47	−40	0.64	No trend
83.2/ Vestlandet	−0.11	−0.24	−21	0.81	No trend
22.22/ Sørlandet	0.35	1.11	93	0.27	No trend

Both aggregations of atmospheric variables and lags between atmospheric variables and streamflow were systematically investigated using a range from two days to three months (88 days) backwards in time. First, an outer loop was used to build n random forest (RF) models, whereby n moving windows of i) cumulative sums and averages and ii) lags were fed to a RF consisting of 25 trees, roughly pruned by limiting the depth of each regression tree to 10. Following this, the weights per model were sorted and the top ten features from $n = 88$ models were fed to a new RF; this inner loop facilitates a fair comparison of the relative importance of different aggregations and lags across the $n \times 10$ highest weighted features. This procedure was followed for each reanalysis dataset.

Table 4
Descriptive statistics of each hydrological station. The lag autocorrelations of 3, 7 and 10 days display the correlated lagged daily mean flows. The 98.5th percentile approximates the flood threshold.

	Minimum flow	Maximum flow	25th percentile	50th percentile	75th percentile	98.5th percentile	3 days lag auto-correlation	7 days lag auto-correlation	10 days lag auto-correlation
Unit	*(m^3/s)*						*(-)*		
234.18/ Finnmark	27.2	3208.0	49.3	92.0	177.3	1144.3	0.88	0.70	0.58
152.4/ Nordland	0.80	290.2	3.8	22.7	49.6	132.7	0.80	0.55	0.44
308.1/ Trøndelag	0.20	188.3	2.3	6.2	15.7	87.2	0.91	0.74	0.64
311.6/ Østlandet	11.9	777.0	25.1	46.8	90.9	289.5	0.95	0.84	0.75
83.2/ Vestlandet	1.7	248.6	13.2	30.7	59.1	144.9	0.86	0.67	0.58
22.22/ Sørlandet	0.04	195.8	3.2	8.8	19.3	76.9	0.69	0.40	0.30

Table 5
Overview of atmospheric and surface variables from reanalysis data used in this study.

Data	Temporal coverage	Spatial resolution	Temporal resolution	Pressure levels (hPa)	Atmospheric variables	Surface variables
ERA5	1979–2018	0.25° x 0.25°	1 h	50; 100; 150; 200; 250; 300; 350; 400; 450; 500; 550; 600; 650; 700; 750; 800; 850; 900; 950; 1000	*RH, T, U, V, W, Z,*	*Cloud, LW, Precip, SLP, SW, Tcv, Tcw, t2m, td2m,*
ERAI	1979–2018	0.75° x 0.75°	6 h			

Table 6
Location of atmospheric columns retrieved from the reanalysis data.

Atmospheric column	*234.18/ Finnmark*	*152.4/ Nordland*	*308.1/ Trøndelag*	*311.6/ Østlandet*	*83.2/ Vestlandet*	*22.22/ Sørlandet*
ERA5	70.00°N, 28.00°E	66.00°N, 13.25°E	64.25°N, 14.00°E	61.25°N, 12.25°E	61.25°N, 6.00°E	58.00°N, 7.75°E
ERAI	69.75°N, 27.75°E	66.00°N, 13.50°E	64.55°N, 14.25°E	61.50°N, 12.00°E	61.50°N, 6.00°E	57.75°N, 7.50°E

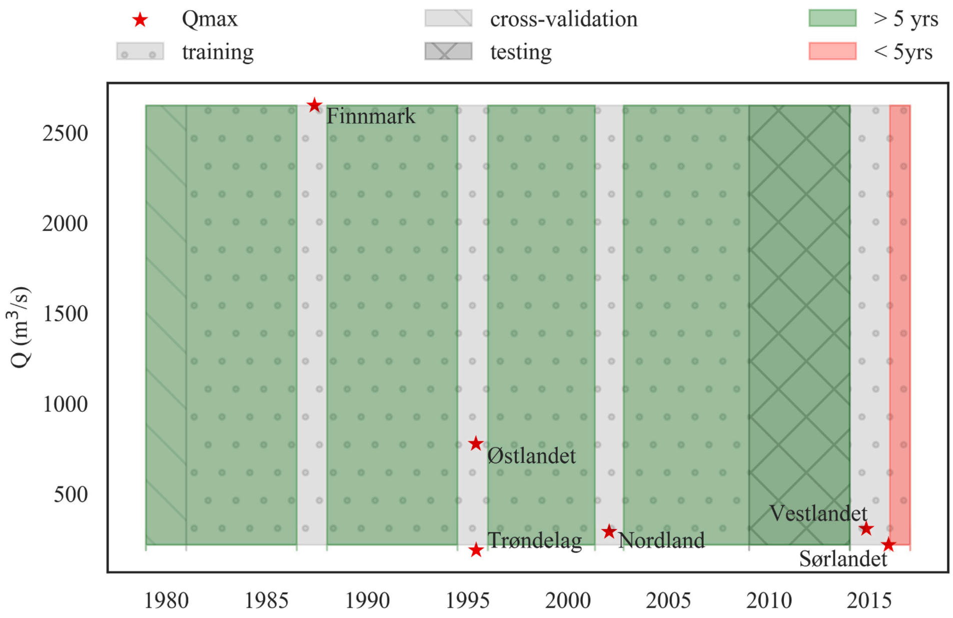


Fig. 3. Splitting of data into training, cross-validation and testing sets.

Table 7 displays the summed weights of the top 50 – 500 highest weighted features extracted from ERA5 for each hydrological station. Corresponding information on automated feature selection from ERAI is given in Table 8. As can be seen, the top 50 highest weighted features account for approximately 80% of the model input across the reanalysis datasets, with a cross-validation coefficient of determination (R^2) ranging from 0.76 to 0.95. Based on this general pattern, the top 50 highest weighted features were extracted as input variables for multiple-linear regression (MLR), random forest (RF), support vector machine (SVM) and multilayer perception (MLP) neural network. To aid visual interpretation, the top 30 highest weighted features extracted from ERA5 and ERAI for each hydrological station (accounting for ~ 80% of summed weights) are shown in Fig. 4. Interestingly, lagged variables led to a far lower out-of-the-bag R^2 than did cumulative sums and averages (not shown). Consequently, only cumulative sums and averages – and not lagged variables – comprise the coarse-resolution and fine-resolution feature selection for each model at each hydrological station.

As seen in Fig. 4, shortwave radiation (*SW*) and temperature variates (*T*) at the surface (*t2m*) and in the boundary layer (1000–850 hPa) are the most important variables, with some variation between ERA5 and ERAI, for the snowmelt-driven flood regime at station 234.18 in Finnmark. The two-month moving window of shortwave radiation from ERAI corresponds to the two-month moving window of temperature at 950 hPa from ERA5 and may be interpreted as reflecting seasonality in the north. The one to three weeks moving windows of surface temperature reflect melting. A similar consistency is seen for the snowmelt-

Table 7
Summed weights of input from ERA5 as a function of the number of features sorted in order of decreasing importance. The corresponding out-of-the-bag coefficient of determination (R^2) on unseen training data is given for 100% of the input.

Number of features in order of decreasing weights	Summed weights (%) from ERA5 variates					
	234.18/ Finnmark	*152.4/ Nordland*	*308.1/ Trøndelag*	*311.6/ Østlandet*	*83.2/ Vestlandet*	*22.22/ Sørlandet*
50	72.6	81.9	76.5	81.4	78.4	66.9
100	82.8	86.5	85.0	86.7	85.2	76.1
150	87.8	89.2	88.6	89.8	88.7	81.1
300	95.0	94.1	94.6	95.1	94.4	90.6
500	98.4	97.6	98.1	98.2	98.0	96.5
Out-of-the-bag R^2	0.92	0.87	0.93	0.95	0.89	0.76

Table 8
Summed weights of input from ERAI as a function of the number of features sorted in order of decreasing importance. The corresponding out-of-the-bag coefficient of determination (R^2) on unseen training data is given for 100% of the input.

Number of features in order of decreasing weights	Summed weights (%) from ERAI variates					
	234.18/ Finnmark	*152.4/ Nordland*	*308.1/ Trøndelag*	*311.6/ Østlandet*	*83.2/ Vestlandet*	*22.22/ Sørlandet*
50	76.4	78.6	76.3	81.2	81.7	71.9
100	84.0	85.1	85.0	87.2	87.1	82.5
150	88.4	88.6	89.1	90.2	90.0	87.3
300	94.9	94.1	94.6	95.2	94.7	93.8
500	98.4	97.8	98.1	98.4	97.9	97.8
Out-of-the-bag R^2	0.92	0.87	0.93	0.95	0.89	0.82

driven flood regime at station 308.1 in Trøndelag and station 311.6 in Østlandet. The persistence of near-surface temperature and vertical wind variates in aggregation windows of one to two months is evident for all stations where snowmelt is the main flood-generating process; this may be interpreted as snow accumulation, for which precipitation (moisture and uplift in the boundary layer) and persistent temperatures around 0 °C are needed. At longer aggregations, mid-level troposphere temperatures and relative humidity are also important. This may link to troposphere-stratosphere interactions and effects on the Jet Stream at 60 N, but conclusive statements upon this cannot be drawn.

For the rainfall-driven flood regime at station 152.4 in Nordland, shorter moving windows of moisture variates, dew point temperature and eastward winds between 700 hPa and 950 hPa are of importance. This distinct difference relative to the snowmelt-driven flood regimes is confirmed by the extracted features for station 83.2 in Vestlandet and 22.22 in Sørlandet. The seasonality in these rainfall-driven flood regimes is represented through moisture and wind variates, as opposed temperature variates in the case of snowmelt-driven flood regimes, reflecting the coastal influence on temperature and the seasonal variation in winds. Moreover, the dominant aggregation period in rainfall-driven flood regimes does not extend beyond roughly three weeks.

In essence, these feature selections reflect that the contribution of snowmelt to streamflow generation requires temperatures around or below the freezing point throughout the winter season (*t2m*), followed by an increase in incoming short-wave radiation (*SW*) and consistent corresponding shifts in the temperature profile in the spring (*T*). The contribution of rainfall to streamflow generation, on the other hand, is linked with the presence and transport (*U*, *V*) of moisture sources (*Tcv*, *Tcw*, *RH*, *td2m*), as well as uplift (*W*) in the boundary layer (1000 – 850 hPa).

The features extracted for station 22.22 in Sørlandet are less consistent between the two reanalysis datasets for periods longer than roughly two weeks. It should be mentioned that where orographic rainfall occurs, the local variability in rainfall can be triggered by sub-grid orographic features. The role of catchment characteristics is only implicitly embedded in the statistical linking of the atmosphere and the hydrological regimes. Sub-grid orographic features will be more pronounced in smaller, steeper catchments, and this in term is less likely to be revealed in the established statistical links at daily level. The feature selection using nested loops of random forests was intentionally mild to minimize the early rejection of possibly useful predictors.

2.3. Model development

2.3.1. Preprocessing of model inputs

The atmospheric data were scaled by removing the median of each feature and dividing by the interquartile range (25th to 75th percentiles). This standardization is more appropriate in the presence of outliers; the resulting input data has zero mean, zero median and a standard deviation of 1, while outliers are preserved without skewing the distribution. Since the hydrological data is skewed, heteroscedastic and strictly positive, a Box-Cox transformation was applied before splitting into training, cross-validation and testing sets.

2.3.2. Baseline

The baseline model in this study is a multiple-linear regression (MLR) model. The preprocessed selected features (top 50 highest weighted features) for each hydrological station subject to transformation were used to construct a MLR model for each of the two reanalysis datasets.

2.3.3. Random forest

A random forest (RF) consists of *n* numbers of regression trees from which the average prediction for a given set of inputs is returned. The regression trees collectively work as an ensemble of piecewise linear models, where instances are sorted until a leaf with a linear regression model is reached. This structure makes RF less sensitive to parameters and hence relatively robust. For each hydrological station, an RF was trained on the preprocessed selected features from the atmospheric column above the relevant station. Pruning was carried out by limiting the depth and increasing the number of samples required for splitting into nodes and leaves with the use of cross-validation. All the RF models were built using bootstrap samples, so that samples, as opposed to the full dataset, of the extracted features were used to train each individual regression tree. The mean-square error was used as the loss function. The models were implemented in Python using *RandomForestRegressor* in *sklearn* (Pedregosa et al., 2011).

2.3.4. Support vector machine

Although originally developed for classification problems, support vector machine (SVM) (Cortes and Vapnik, 1995) is also applicable to regression problems. The idea behind this algorithm is to maximize the margins to an *n-1* dimensional hyperplane for an *n*-dimensional dataset, so that the distance between the closest vectors – or support vectors – and the hyperplane is maximized. To maximize this distance, a loss function is used, whereby smaller distances between the support vectors and the hyperplane are costlier than larger distances. For regression problems, this hyperplane is a regression curve defined by the support vectors. Kernels such as a radial basis function may be used to handle non-linearity if data is not linearly separable after standardization and mapping to a higher dimensional space. In essence, the kernel then behaves as a weighted *k* nearest neighbor algorithm for unseen data in infinite dimensions, in which closer observations have a larger influence on the prediction of unseen data.

For each hydrological station, an SVM was trained on the preprocessed selected features from the atmospheric column above the relevant station. Two parameters in the SVM model were tuned with the use of cross-validation. These parameters respectively control i) the strength of L2 regularization, and ii) the epsilon-tube, denoting the difference between the predicted and observed value for which no penalty occurs during training. The SVM models were implemented with a radial basis function in Python using *SVR* in *sklearn* (Pedregosa et al., 2011).

2.3.5. Multilayer perceptron

Multilayer perceptron (MLP) is a feed-forward neural network. Inputs propagate through the hidden layers by means of activation functions. Each hidden node is associated with a bias and weights per input,

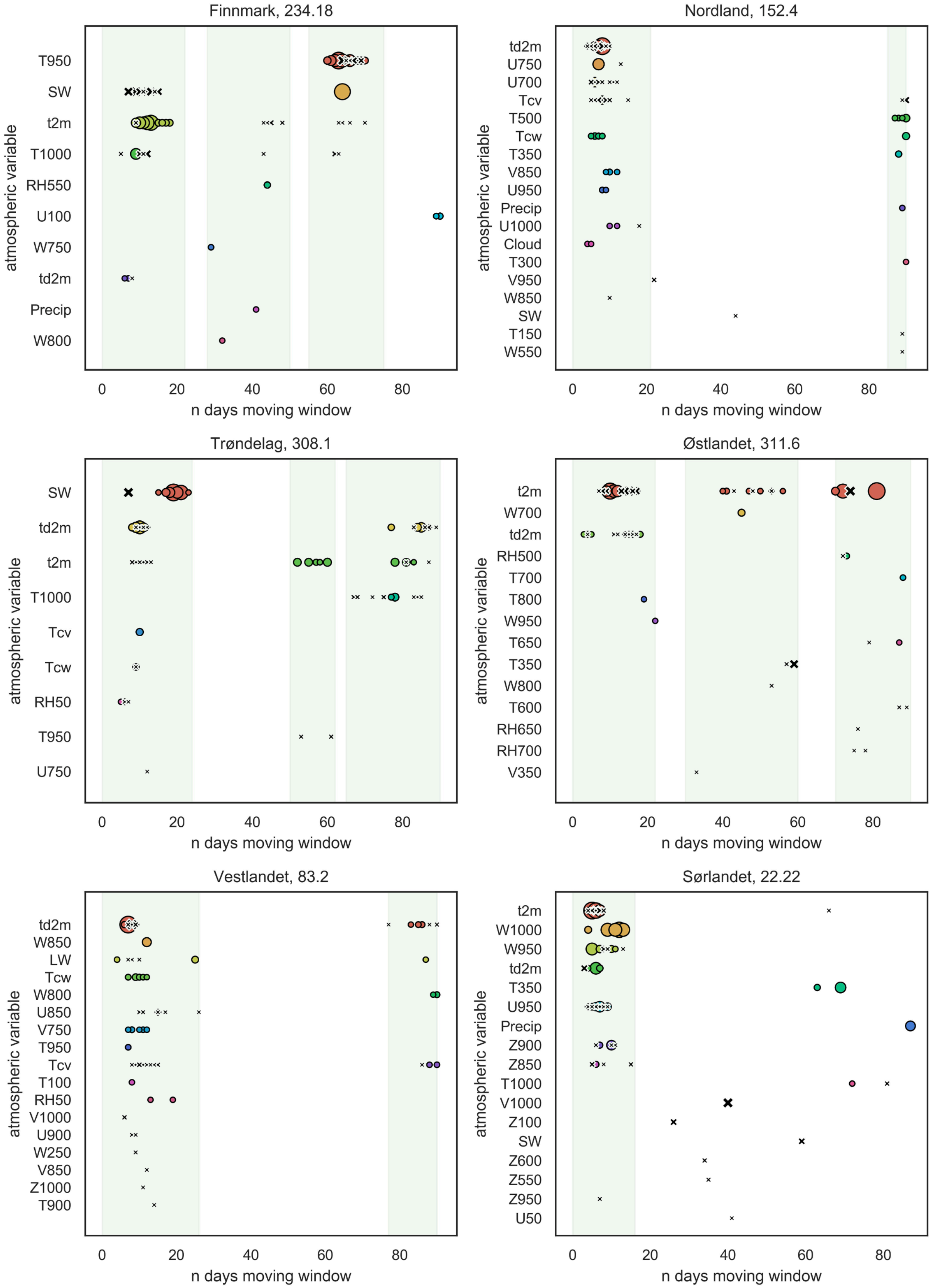


Fig. 4. Selected features from ERA5 (black crosses) and ERAI (colored markers) in order of decreasing importance (see Table A1 in Appendix A for list of symbols). The relative importance of features is reflected by marker size. Shaded areas show pronounced aggregation windows.

and the activation function transfers the signal until the output layer is reached. To avoid saturation of neurons, the inputs should be standardized and normalized. To avoid overfitting, regularization methods such as dropout or weight decay can be applied.

For each hydrological station, an MLP was trained on the preprocessed selected features using stochastic gradient decent. Hyperbolic tangent activation functions and rectified linear units were used in the hidden layers along with L2 regularization. MLP differs from RF and SVM by being more complex, less transparent due to randomized noise in the hidden layers, and more sensitive to parameters. Parameters were tuned using the cross-validation set. In addition to the use of cross-validation for parameter tuning, early stopping was employed; when the loss on an arbitrary 10% of the training data stopped decreasing above a threshold value (0.0005–0.5), training was stopped. With one to three hidden layers per model, the MLP models constructed may be referred to as shallow neural networks. The models were implemented in Python using *MLPRegressor* in *sklearn* (Pedregosa et al., 2011).

2.3.6. Model complexity

In the context of machine learning, the term “model complexity” is often used to reflect the dimensionality of the space of possible models, expressed as the number of degrees of freedom measured by the adjustable hyper-parameters. As each of the models described above was developed for each station and dataset with a unique set of hyper-parameters and inputs, a conceptual classification of relative model complexity based on general model characteristics shared between models employing the same algorithm was used to assess the improvements to accuracy with increasing model complexity (see Table 9). This allows for a separation between model performance resulting from strong and near-linear coupling between the large-scale atmospheric circulation and the catchment (small inter-model differences) and model performance resulting from the structural complexity in which non-linearity is handled (consistently larger inter-model differences). The word “reconstruct” is used to refer to this translation from large-scale atmospheric circulation to streamflow using major identified drivers only; it should not be confused with prediction, in which catchment characteristics will also contribute as explicit model inputs. An illustrative flowchart of the methodology is provided in Fig. 5.

Table 9
Overview of model characteristics and complexity classification.

Model	*Model type*	*Model structure*	*Hyper-parameters*	*Model complexity (discretized)*
MLR	Parametric	Linear	–	Naïve(0)
RF	Non-parametric	Piece-wise linear ensemble averaging	Number of trees, pruning criteria (e.g. maximum depth or number of samples per node or leaf)	Simple(1)
SVR	Non-parametric	Nonlinear kernel	Epsilon-tube, L2-penalty, tolerance	Intermediate (2)
MLP	Non-parametric	Nonlinear hidden layers	Number of hidden layers, activation function, solver, L2-penalty, learning rate, momentum, number of iterations, tolerance	Complex(3)

2.4. Evaluation

Four metrics commonly applied for evaluation of hydrological models were used to assess the performance of the machine learning models*: Pearson’s r* (R) (Equation (1)), *normalized root-mean-square error* (NRMSE) (Equation (2)), *Nash-Sutcliffe Efficiency* (NSE) (Equation (3)), and *Kling-Gupta Efficiency* (KGE) (Equation (4)). *Pearson’s r* measures the strength of a positive or negative linear relationship between the reconstructed and observed flows. However, it does not distinguish systematic errors; therefore, the root-mean-square error normalized by mean flow was used to indicate discrepancies in flow magnitudes between reconstructed and observed flows. *Nash-Sutcliffe Efficiency* (Nash and Sutcliffe, 1970) is useful to assess model performance in relation to a ‘mean-flow benchmark’, whereby positive values indicate accuracy above estimating a particular value in the data series as the mean of all data points. Lastly, the *Kling-Gupta Efficiency* is a decomposition of the NSE into correlation, variability bias and mean bias and is intended to improve certain shortcomings (Gupta et al., 2009); hence, KGE has gained popularity for model evaluation in the hydrological community. However, KGE does not have the intuitive mean flow benchmark (NSE = 0). In fact, using the mean flow as benchmark predictor results in $KGE = 1 - \sqrt{2} \approx -0.41$ (Knoben et al., 2019). Therefore, KGE cannot be regarded a substitute for NSE nor can it be directly compared; rather all aforementioned components of KGE should be considered, and a concurrent use of KGE and NSE may constitute a better foundation for model evaluation. In the equations below, n denotes the sample size, *Qrec* is the reconstructed flow, *Qobs* is the observed flow, R^2 is the coefficient of determination, σ is the standard deviation, and overbars indicate arithmetic means. Box-Cox back-transformation was performed before the models were assessed with R, NRMSE, NSE and KGE.

$$R = \frac{\sum_{i=1}^{n}(Qobs_i)(Qrec_i)}{\sqrt{\sum_{i=1}^{n}(Qobs_i)^2\sum_{i=1}^{n}(Qrec_i)^2}} \tag{1}$$

$$NRMSE = \sqrt{\frac{\sum_{i=1}^{n}(Qrec_i - Qobs_i)^2}{n}} * \frac{1}{\overline{Qobs}} \tag{2}$$

$$NSE = 1 - \frac{\sum_{i=1}^{n}(Qrec_i - Qobs_i)^2}{\sum_{i=1}^{n}\left(Qobs_i - \overline{Qobs}\right)^2} \tag{3}$$

$$KGE = 1 - \sqrt{(R^2 - 1)^2 + \left(\frac{\sigma_{Qprec}}{\sigma_{Qobs}} - 1\right)^2 + \left(\frac{\overline{Qrec}}{\overline{Qobs}} - 1\right)^2} \tag{4}$$

3. Results

Fig. 6 shows the flow duration curves of observed and reconstructed flows in the testing period (2009–2014). Time series of the corresponding daily flows are shown for multiple linear regression (MLR), random forest (RF), support vector machine (SVM), and multilayer perceptron (MLP) separately in Fig. 7, Fig. 8, Fig. 9 and Fig. 10 respectively. At first glance, the flow duration curves seem similar. However, flows below exceedance probabilities of 10% and above 90% are consistently smaller and larger than the observed flows for MLR. This is not observed for the machine learning models. A closer look at the time series of daily streamflow as reconstructed from the identified major drivers reveal that MLR fails to capture peaks and only gives a rough representation of the seasonal variation in streamflow.

While the baseline model, MLR, clearly is outperformed by the machine learning models, both at high flows and low flows, some differences in relation to model complexity and hydrological regimes are also evident. The simpler model, RF, captures the seasonal behavior in snowmelt-driven flood regimes (234.18, 308.1 and 311.6), but tends to underestimate the magnitude of the annual peaks. SVM, on the other hand, captures the timing of these peaks better, but does so at the expense of some overestimation, particularly the annual peak in 2010 in Finnmark. The most accurate and timely reconstruction of streamflow that clearly distinguishes between high flows and low flows is obtained with MLP.

RF also fails to capture the magnitude of the larger peaks at stations 152.4, 83.2 and 22.22. Comparing the peaks in these rainfall-driven

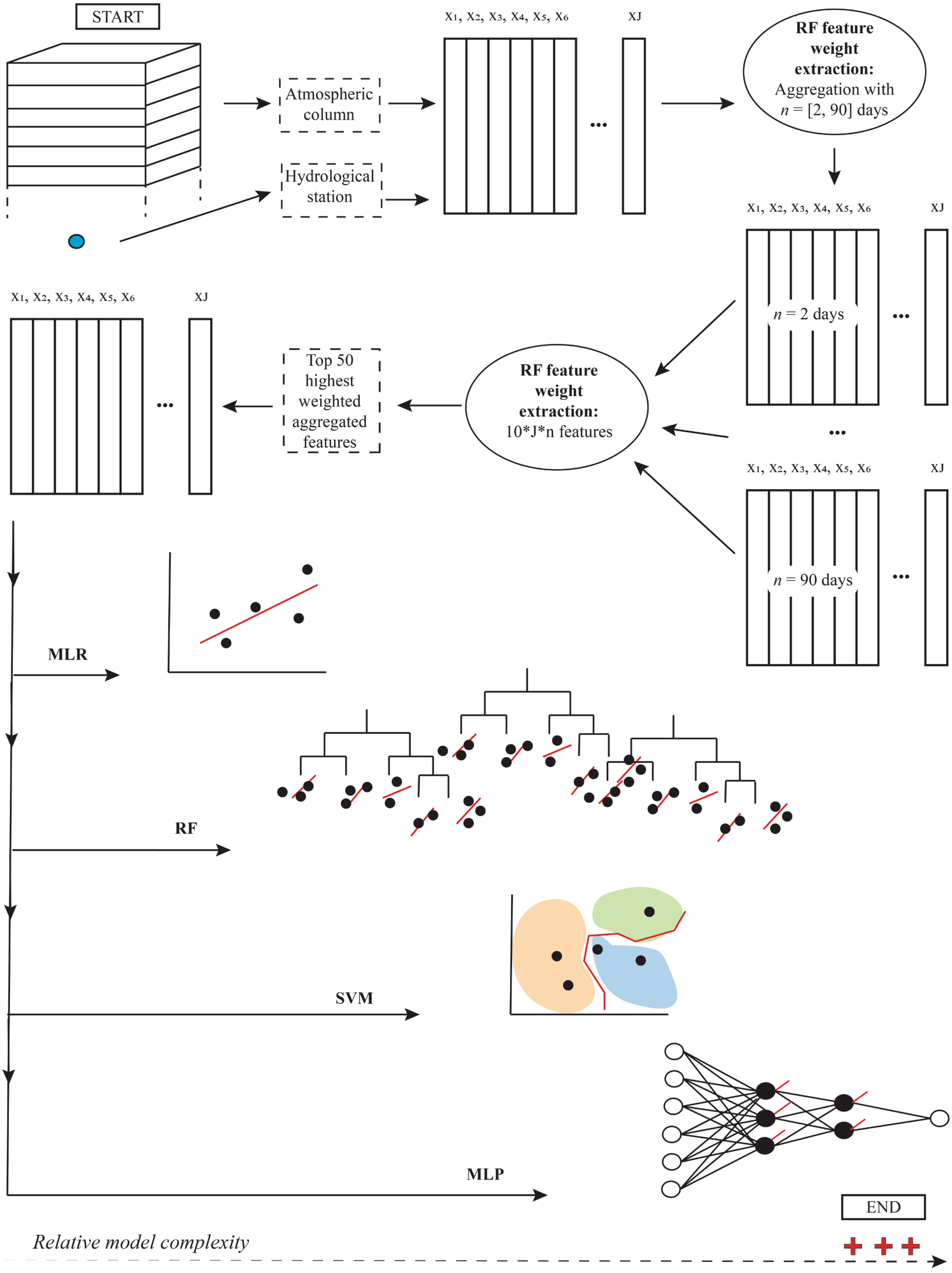

START
Atmospheric column
Hydrological station
X1, X2, X3, X4, X5, X6
XJ
RF feature weight extraction: Aggregation with n = [2, 90] days
n = 2 days
n = 90 days
RF feature weight extraction: 10*J*n features
Top 50 highest weighted aggregated features
MLR
RF
SVM
MLP
END
Relative model complexity
+ + +

Fig. 5. Workflow of feature selection and model development.

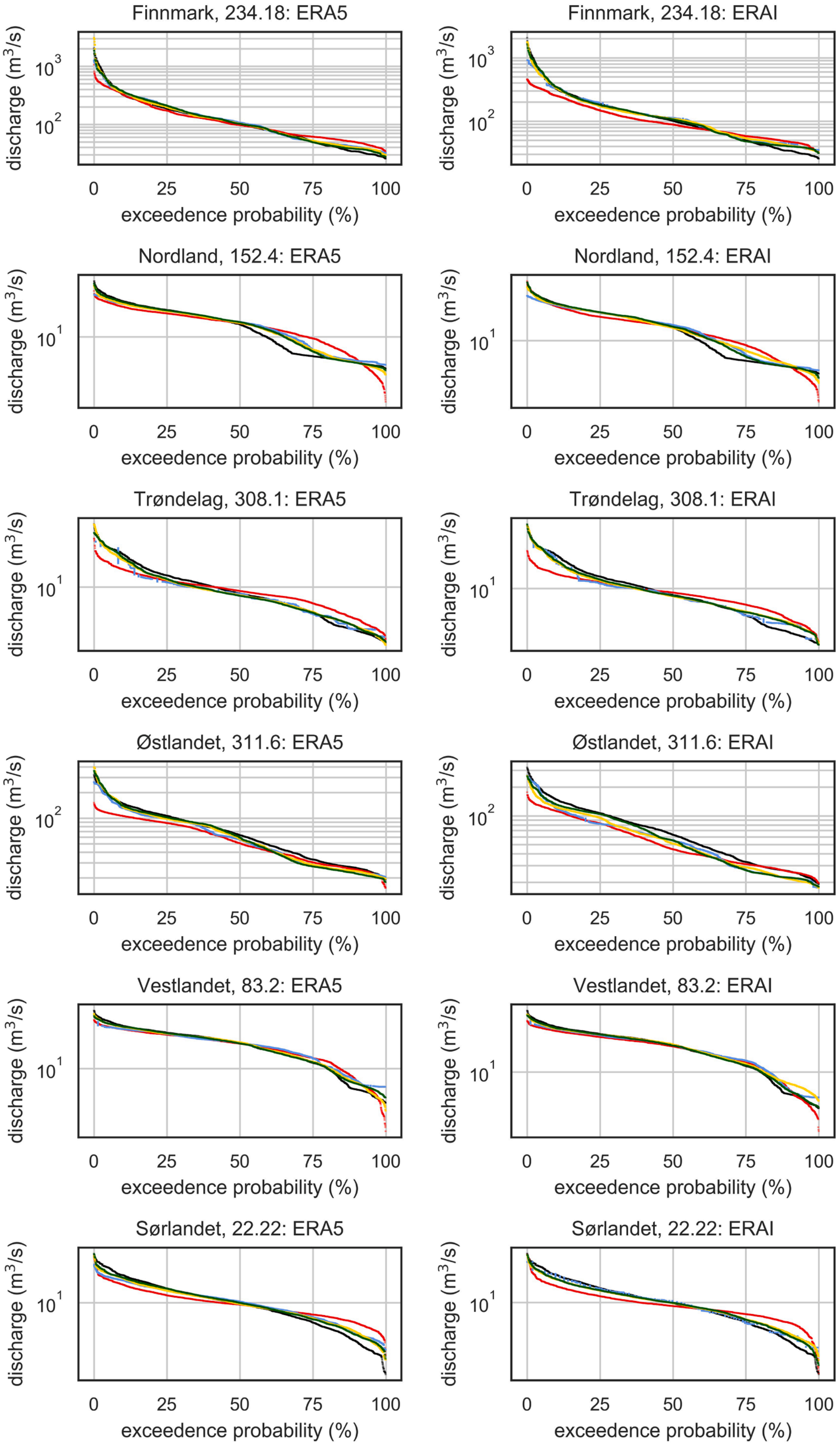


Fig. 6. Flow duration curves of observed (black) and predicted flows by MLR (red), RF (blue), SVM (yellow) and MLP (green) using extracted features from ERA5 (left) and ERAI (right). (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

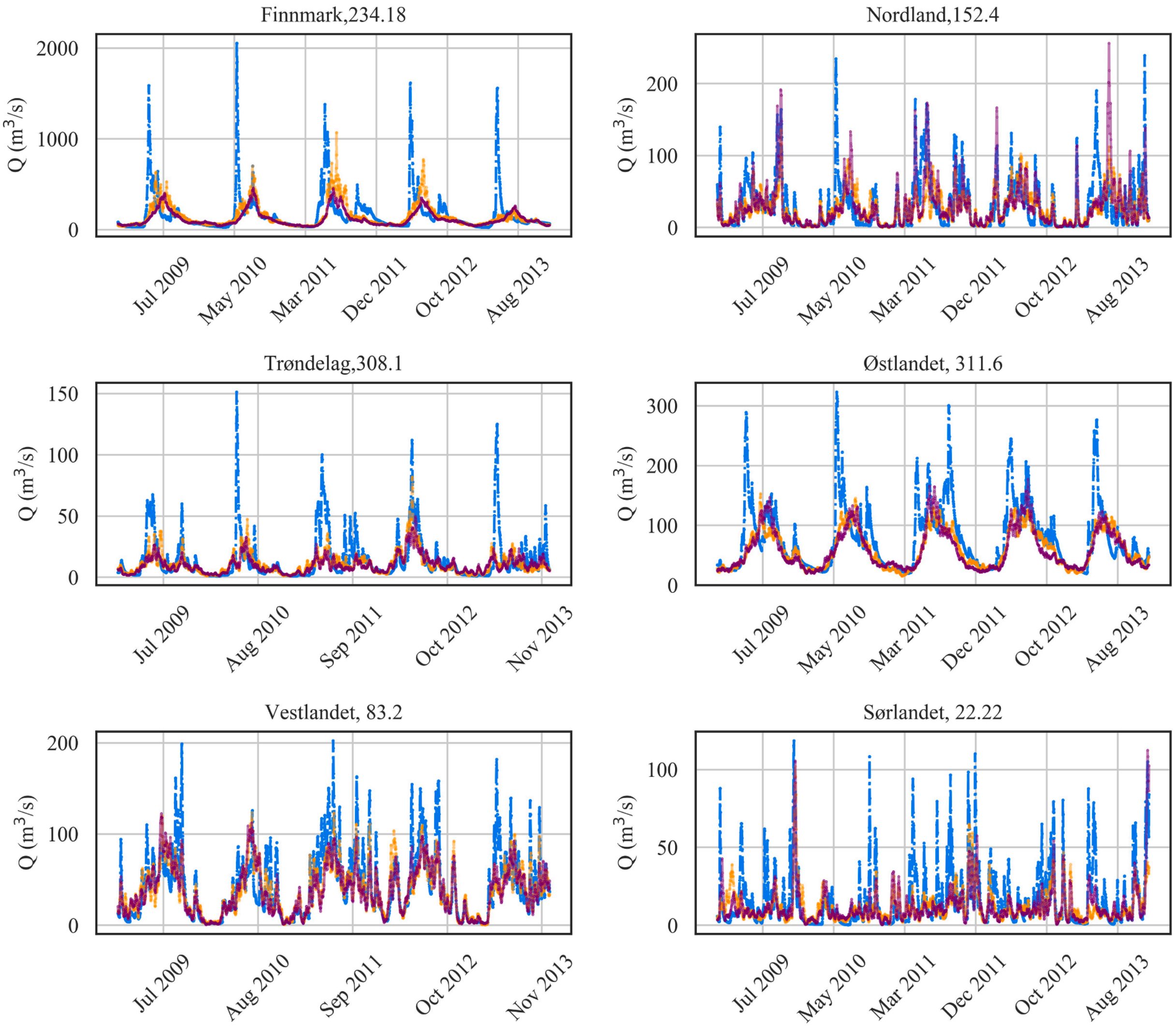


Fig. 7. Observed (blue) and reconstructed flows from identified drivers by baseline model (MLR) using data from ERA5 (orange) and ERAI (purple). (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

flood regimes as reconstructed from the identified major drivers by RF, SVM and MLP, it is evident that increasing model complexity leads to more distinct peaks. Furthermore, a more accurate separation of high flows and low flows is obtained with MLP.

For the identification of major large-scale atmospheric drivers of daily streamflow, ERA5 is not consistently favored over ERAI. However, effects related to the scale of the atmospheric input relative to the size of a given catchment was not specifically investigated in this analysis. Nevertheless, differences in reconstructed flows for each station may be viewed in relation to the consistency between identified drivers from the two reanalysis datasets. This is most apparent for station 22.22, where identified drivers beyond two weeks are scattered in space and time across ERA5 and ERAI (see Fig. 4, Section 2.2). In the case of station 311.6, where mixed processes also contribute to the generation of high flows, the reconstructed streamflow is consistently higher with ERA5, and the inter-model differences between RF, SVM and MLP are generally smaller than for the other stations.

Metrics for training and testing of MLR, RF, SVM and MLP are given per station in Tables 10 – 15. The best scores on the testing sets are highlighted in bold. The poorest performance is seen for station 22.22; only MLP captures some of the streamflow dynamics at this station – and to a larger extent with the coarser reanalysis data (ERAI).

Scatter plots comparing the performance of RF, SVM and MLP against the baseline model on the testing set are provided in Fig. 11. The corresponding score ranges obtained using identified drivers of daily streamflow are displayed for each metric in Fig. 12. In summary, the highest performance in terms of correlation (R), normalized root-mean-square error (NRMSE), Nash-Sutcliffe Efficiency (NSE) and Kling-Gupta Efficiency (KGE) is obtained with MLP. Furthermore, the lowest drop on the aforementioned metrics from training to testing is also obtained with MLP. It should be emphasized that the scores obtained with simpler metrics (R and NRMSE) on the testing set at times differ marginally between the machine learning models; however, when summarizing the highest scores on the more sophisticated metrics (NSE and KGE) from Tables 10-15, the more complex models (SVM and MLP) perform consistently better.

4. Discussion

4.1. Model performance

The direct translation from identified major drivers to daily streamflow was most accurately made with the most complex model (MLP), with a Nash-Sutcliffe Efficiency (NSE) on the testing set ranging from 0.71 to 0.81 on all stations except for station 22.22, representing the smallest catchment. Although all models performed substantially worse on station 22.22, MLP still obtained the highest NSE there as well (0.59). Patterns of increasing accuracy with model complexity were

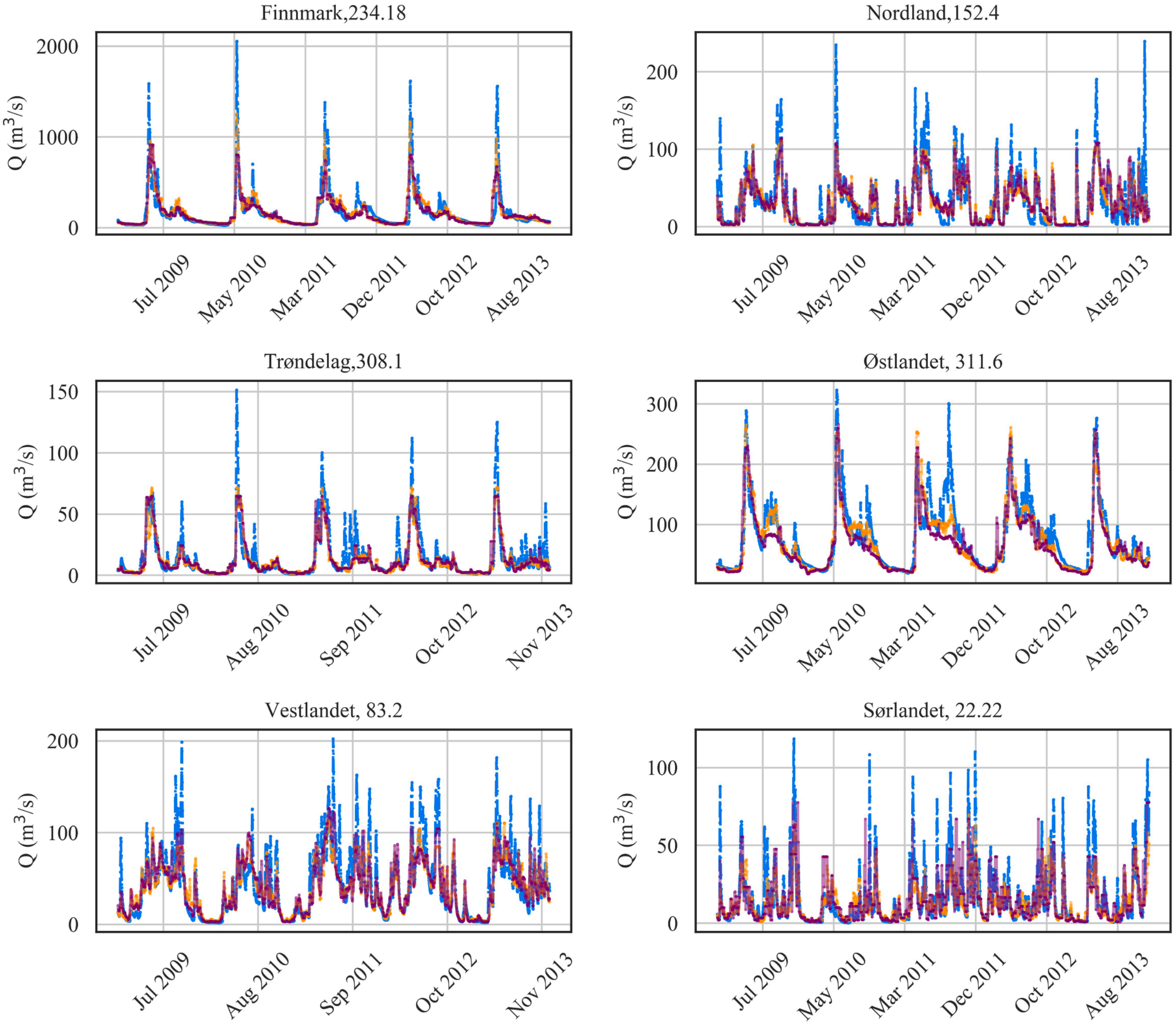


Fig. 8. Observed (blue) and reconstructed flows from identified drivers by RF using data from ERA5 (orange) and ERAI (purple). (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

found, indicating that structural complexity is needed to exploit information in the identified major drivers – although the machine learning models showed marginal differences on simpler metrics like correlation. From the results, it is clear that a simple multiple-linear regression (MLR) is unsuitable for direct downscaling from large-scale atmospheric variables to daily streamflow. This is in alignment with previous findings (Cannon and Whitfield, 2002) and does not require further elaboration here.

The drops in NSE from training to testing provide an indication of the robustness of the machine learning models, MLP had an average drop of 0.03 with ERA5 and less than 0.01 with ERAI. A slightly larger drop in NSE (0.04 – 0.05) was observed for SVM and the largest drop (0.11) was observed for RF, mostly due to the inability to generalize for station 22.22. These drops inversely reflect the robustness of the models. In other words, the most accurate model, MLP, is also the most robust model. This may seem counter-intuitive, as MLP is more sensitive to parameters than RF, but nevertheless supports conclusions from previous studies, showing that there are limitations to linear models when applied to nonlinear systems with complex interactions between inputs and outputs (Cannon and Whitfield, 2002). In essence, a shallow neural network may generalize well when parameters are tuned optimal or near-optimal with the use of cross-validation. The fact that the drop in NSE was somewhat larger with the fine-resolution reanalysis data (0.25°) may be an artifact of scale issues. Nogueira (2020) assessed the differences in spatiotemporal resolution and convective and microphysical parameterizations between ERA5 and ERAI by relating systematic and random components of the differences in rainfall against the differences in temperature, water vapor, evaporation, moisture flux divergence and pressure vertical velocity. The study concluded that the spatial and temporal representation of rainfall, as well as other variables, in ERA5 and ERAI have similar error and bias except in the tropics, where ERA5 provides improvements due to the increased quantity and quality of assimilated observations. In other words, ERAI may be preferred to ERA5 for the specific use and purpose of this study when only a single atmospheric column is to be considered.

The substantially lower model performance at station 22.22 in Sørlandet must be viewed from a catchment perspective. Station 22.22 is located in a catchment that differs from the other catchments in a number of ways. Firstly, it is the smallest catchment of the six considered, has the second largest annual flow normalized by catchment area and the lowest lag autocorrelation as compared to the other stations. This indicates a rapidly responding system. Needless to say, the role of atmospheric forcing on finer spatiotemporal scales may therefore be of greater importance here than for the other stations. Secondly, it has the most dominantly rainfall-driven flood regime. Sivakumar et al. (2001) identified the possible existence of chaos in rainfall-runoff processes; this in turn complicates streamflow reconstruction in rainfall-driven flood regimes. Thirdly, both feature selections from ERA5 and ERAI

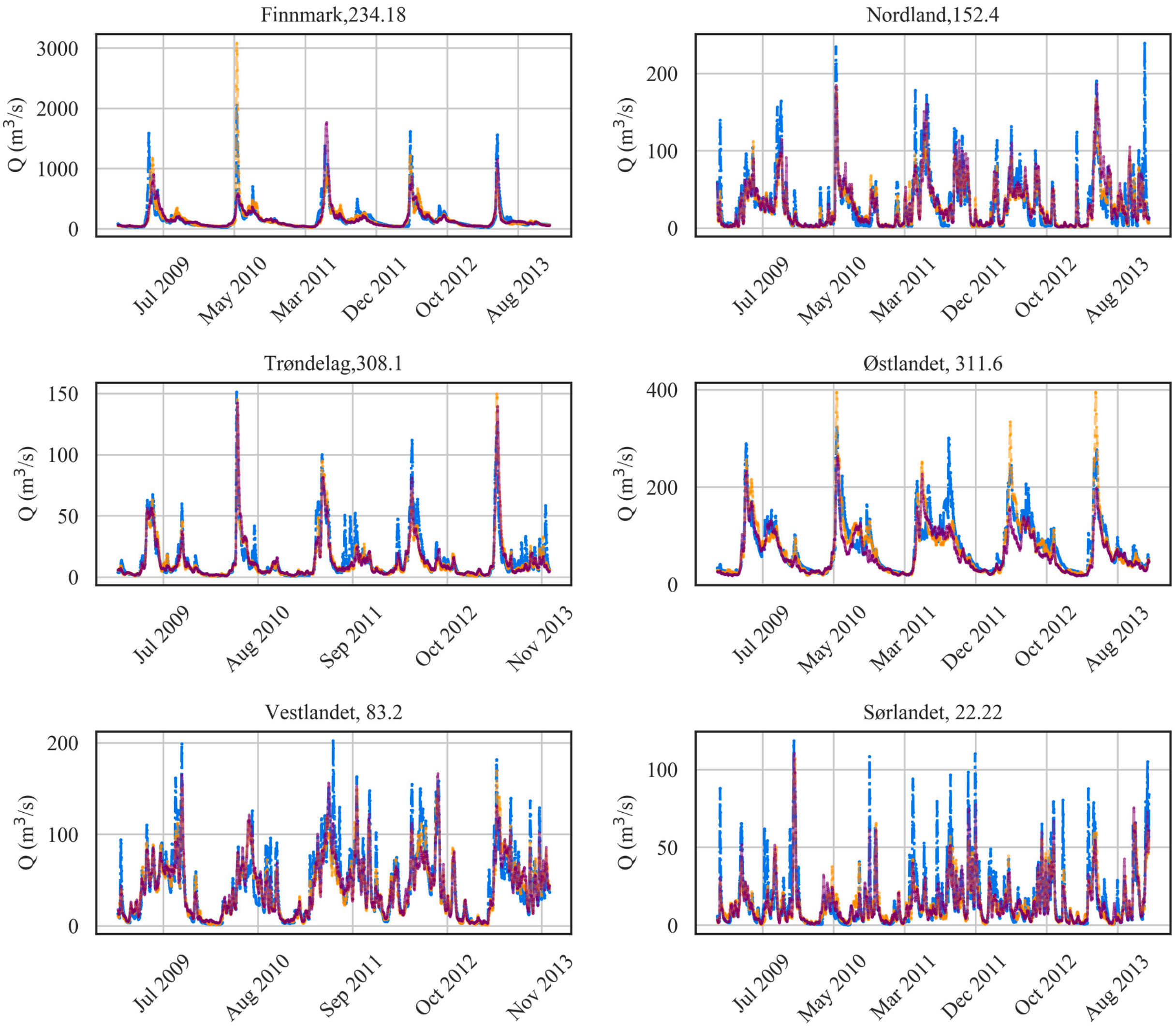


Fig. 9. Observed (blue) and reconstructed flows from identified drivers by SVM using data from ERA5 (orange) and ERAI (purple). (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

show the lowest dominant aggregation period for this station, with the largest disagreement on aggregated atmospheric variables beyond roughly two weeks. This indicates that coupling between the large-scale atmospheric circulation and the catchment not only occurs on a shorter time-scales, but also that catchment characteristics, like slope, permeability, land use and antecedent soil moisture; sub-grid features, like topography; and local variability may play an equally, or more, important role in streamflow generation at the temporal resolution of one day. A traditional hydrological modelling approach could be used to investigate the relative importance of local variability, sub-grid topography and catchment characteristics in depth.

For prediction tasks, catchment characteristics must undoubtedly be included for all stations; however, the relative improvements in accuracy with the inclusion of catchment characteristics can be expected to be higher for station 22.22. This, in turn, may from a hydrological climate-impact perspective, point to the potential of direct downscaling with shallow neural networks in catchments where forcing from large-scale atmospheric circulation dominates local variability, sub-grid features and catchment characteristics in daily streamflow generation to the degree that an NSE in the range of 0.7 – 0.8 can be obtained on daily testing data spanning five consecutive years. As such, the methodology presented in this paper may be used to filter out suitable catchments for direct downscaling in the context of hydrological climate-impact modelling. Among the six catchments considered, the smallest inter-model difference in performance was observed for station 311.6 in Østlandet, while the overall intra-model performance was consistent for all but station 22.22 in Sørlandet. Station 311.6 would therefore be most suitable for direct downscaling, while station 22.22 would not be suitable without complementary modelling using more physically-based approaches.

The inability to reconstruct the magnitude of annual peak flows accurately, with a tendency towards underestimation by RF, overestimation by SVM and a modest combination of both by MLP, is partially a reflection of the loss functions used to optimize the models (mean-flow metrics) and may hence improve with customized loss functions particularly penalizing error on high flows. While a complex model structure should not be used to compensate for an inadequate selection of input variables, a thorough investigation with deep learning should be undertaken, in which variates selected based on i) expert knowledge/ prior assumptions, ii) identified major drivers from an automated feature selection procedure as presented in this paper and iii) all available data respectively are investigated in relation to increasing model complexity. Additionally, some catchments are influenced by complex catchment characteristics that lag or amplify the catchment response dynamically in space and time; for such catchments, complex statistics are required to infer a forcing-response relationship from the atmosphere, and the direct linking between large-scale atmospheric circulation and daily streamflow with traditional machine learning

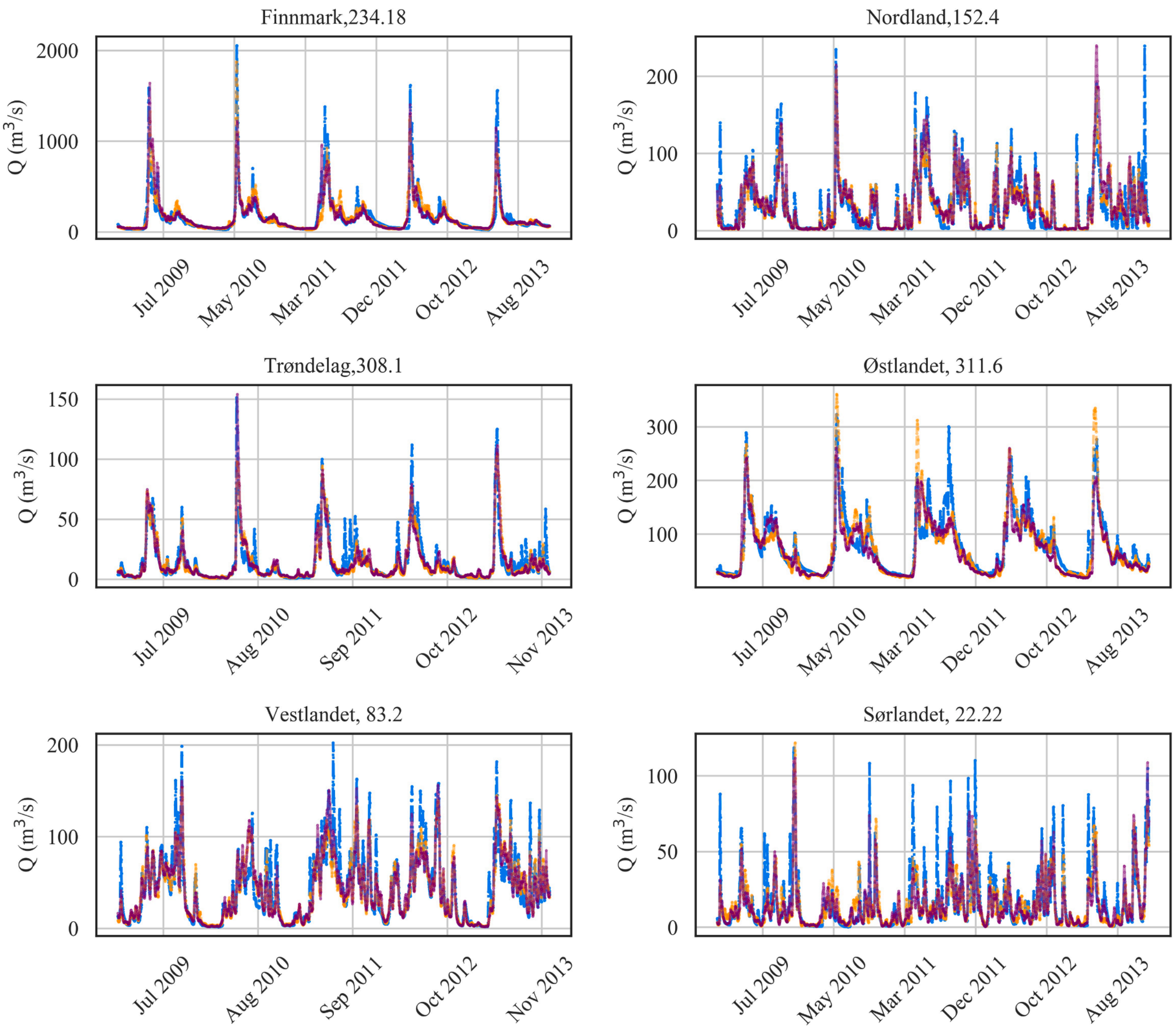


Fig. 10. Observed (blue) and reconstructed flows from identified drivers by MLP using data from ERA5 (orange) and ERAI (purple). (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

Table 10
Training and testing metrics for station 234.18 in Finnmark.

Model	R				NRMSE				NSE				KGE			
	ERA5		*ERAI*		*ERA5*		*ERAI*		*ERA5*		*ERAI*		*ERA5*		*ERAI*	
	train	*test*	*train*	*test*	*train*	*test*	*train*	*test*	*train*	*test*	*train*	*test*	*train*	*test*	*train*	*test*
MLR	0.37	0.32	0.42	0.29	1.42	1.29	1.37	1.29	−1.83	−1.86	−4.34	−5.44	−0.20	−0.20	−0.64	−0.77
RF	0.92	0.86	0.88	0.83	0.65	0.68	0.77	0.75	0.63	0.59	0.42	0.38	0.61	0.72	0.50	0.57
SVM	0.88	0.84	0.79	0.79	0.71	0.78	0.93	0.82	0.70	0.69	0.53	0.52	0.81	0.82	0.76	0.76
MLP	0.90	**0.87**	0.89	**0.89**	0.63	**0.64**	0.67	**0.61**	0.75	**0.71**	0.73	**0.75**	0.81	**0.83**	0.82	**0.86**

Table 11
Training and testing metrics for station 152.4 in Nordland.

Model	R				NRMSE				NSE				KGE			
	ERA5		*ERAI*		*ERA5*		*ERAI*		*ERA5*		*ERAI*		*ERA5*		*ERAI*	
	train	*test*	*train*	*test*	*train*	*test*	*train*	*test*	*train*	*test*	*train*	*test*	*train*	*test*	*train*	*test*
MLR	0.65	0.57	0.71	0.67	0.88	0.95	0.84	0.89	−0.37	−0.87	0.37	0.23	0.35	0.20	0.67	0.63
RF	0.87	0.84	0.88	0.84	0.57	0.64	0.57	0.64	0.55	0.41	0.57	0.46	0.64	0.59	0.65	0.63
SVM	0.89	0.86	0.88	0.83	0.55	0.60	0.54	0.64	0.62	0.57	0.68	0.61	0.68	0.71	0.76	0.79
MLP	0.90	**0.88**	0.90	**0.86**	0.49	**0.55**	0.48	**0.60**	0.75	**0.71**	0.76	**0.73**	0.82	**0.82**	0.83	**0.86**

Table 12
Training and testing metrics for station 308.1 in Trøndelag.

Model	R				NRMSE				NSE				KGE			
	ERA5		*ERAI*		*ERA5*		*ERAI*		*ERA5*		*ERAI*		*ERA5*		*ERAI*	
	train	*test*	*train*	*test*	*train*	*test*	*train*	*test*	*train*	*test*	*train*	*test*	*train*	*test*	*train*	*test*
MLR	0.55	0.52	0.56	0.51	1.16	1.16	1.16	1.20	−2.82	−2.76	−3.83	.-5.49	−0.39	−0.35	−0.63	−0.93
RF	0.88	0.88	0.88	0.86	0.66	0.68	0.67	0.69	0.55	0.55	0.55	0.56	0.60	0.60	0.61	0.65
SVM	0.91	0.88	0.90	0.88	0.57	0.64	0.59	0.63	0.73	0.74	0.74	0.75	0.75	0.83	0.78	0.83
MLP	0.92	**0.90**	0.92	**0.90**	0.52	**0.57**	0.54	**0.57**	0.80	**0.78**	0.79	**0.79**	0.82	**0.85**	0.82	**0.87**

Table 13
Training and testing metrics for station 311.6 in Østlandet.

Model	R				NRMSE				NSE				KGE			
	ERA5		*ERAI*		*ERA5*		*ERAI*		*ERA5*		*ERAI*		*ERA5*		*ERAI*	
	train	*test*	*train*	*test*	*train*	*test*	*train*	*test*	*train*	*test*	*train*	*test*	*train*	*test*	*train*	*test*
MLR	0.54	0.68	0.54	0.65	0.71	0.56	0.71	0.59	−2.25	−0.90	−1.92	−0.82	−0.20	0.17	−0.10	0.22
RF	0.93	**0.90**	0.90	0.86	0.32	**0.32**	0.37	0.38	0.77	0.76	0.68	0.62	0.76	0.84	0.72	0.74
SVM	0.92	0.88	0.90	0.88	0.33	0.36	0.38	0.38	0.78	0.77	0.66	0.56	0.80	0.85	0.69	0.67
MLP	0.93	**0.90**	0.92	**0.91**	0.30	0.33	0.34	**0.31**	0.83	**0.81**	0.77	**0.76**	0.85	**0.87**	0.80	**0.82**

Table 14
Training and testing metrics for station 83.2 in Vestlandet.

Model	R				NRMSE				NSE				KGE			
	ERA5		*ERAI*		*ERA5*		*ERAI*		*ERA5*		*ERAI*		*ERA5*		*ERAI*	
	train	*test*	*train*	*test*	*train*	*test*	*train*	*test*	*train*	*test*	*train*	*test*	*train*	*test*	*train*	*test*
MLR	0.75	0.78	0.75	0.77	0.56	0.53	0.56	0.55	0.11	0.24	0.09	0.16	0.51	0.56	0.50	0.52
RF	0.85	0.85	0.88	0.86	0.46	0.46	0.42	0.43	0.45	0.44	0.59	0.56	0.61	0.60	0.68	0.69
SVM	0.91	**0.90**	0.91	0.89	0.35	**0.37**	0.35	0.38	0.76	0.75	0.75	0.74	0.79	**0.84**	0.77	**0.83**
MLP	0.91	**0.90**	0.91	**0.90**	0.36	**0.37**	0.35	**0.36**	0.76	**0.76**	0.78	**0.78**	0.81	**0.84**	0.82	**0.86**

Table 15
Training and testing metrics for station 22.22 in Sørlandet.

Model	R				NRMSE				NSE				KGE			
	ERA5		*ERAI*		*ERA5*		*ERAI*		*ERA5*		*ERAI*		*ERA5*		*ERAI*	
	train	*test*	*train*	*test*	*train*	*test*	*train*	*test*	*train*	*test*	*train*	*test*	*train*	*test*	*train*	*test*
MLR	0.57	0.56	0.63	0.67	1.00	0.99	0.97	0.90	−0.90	−1.25	−0.04	−0.49	0.18	0.07	0.49	0.25
RF	0.86	0.77	0.84	0.79	0.66	0.77	0.65	0.71	0.32	−0.15	0.53	0.43	0.47	0.32	0.68	0.68
SVM	0.83	0.78	0.88	**0.86**	0.69	0.73	0.58	**0.60**	0.35	0.32	0.62	0.56	0.55	0.60	0.69	0.67
MLP	0.85	**0.81**	0.87	0.85	0.63	**0.68**	0.61	0.61	0.55	**0.50**	0.60	**0.59**	0.68	**0.72**	0.71	**0.72**

techniques is unsuitable. To separate atmospheric forcing from anthropogenic forcing, the selected catchments should not be heavily developed, regulated or otherwise strongly influenced by human activity.

4.2. Interpretability and physical consistency

The automated feature selection procedure presented in this paper differs from methods used in previous studies (see Table 1, Section 1). In contrast with [canonical] correlation analysis, stepwise predictor selection, maximal information coefficient or simply prior assumptions, this study has used the weights assigned by roughly pruned, bootstrapped random forests to extract relevant features. Furthermore, dimensionality reduction was not carried out; rather, the resulting feature selections comprised aggregated variables from two-daily to three-monthly moving windows. Consequently, the feature selections retain both cross-correlations and autocorrelations within the data, and this translates into information in the most complex model, MLP. The setup using two different atmospheric reanalysis datasets provides a validation of the feature selection procedure. Although some variability was seen between the extracted features from ERA5 and ERAI, the general patterns of the extracted temperature variates and shortwave radiation for station 234.18 in Finnmark, station 308.1 in Trøndelag and station 311.6 in Østlandet were consistent and interpretable for snowmelt-driven flood regimes. Likewise, the importance of shorter moving windows of moisture variates, dew point temperature and boundary layer winds was consistent for the rainfall-driven flood regimes at station 152.4 in Nordland, station 83.2 in Vestlandet and station 22.22 in Sørlandet. Seasonality was represented with aggregations of temperature variates for snowmelt-driven flood regimes and moisture and wind variates for rainfall-driven flood regimes at two to three months.

The inconsistency in features extracted for station 22.22 in Sørlandet beyond two weeks may, as mentioned above, indicate the relative importance of catchment characteristics or the inability of the selected traditional machine learning techniques to treat complex statistics. Furthermore, all atmospheric reanalysis datasets contain biases. When biases across two reanalysis datasets differ, the resulting feature selections will differ accordingly. Moreover, the feature selection procedure with looped random forests assumed the same pruning depth for all regression trees and a constant retrieval of ten candidate variables per

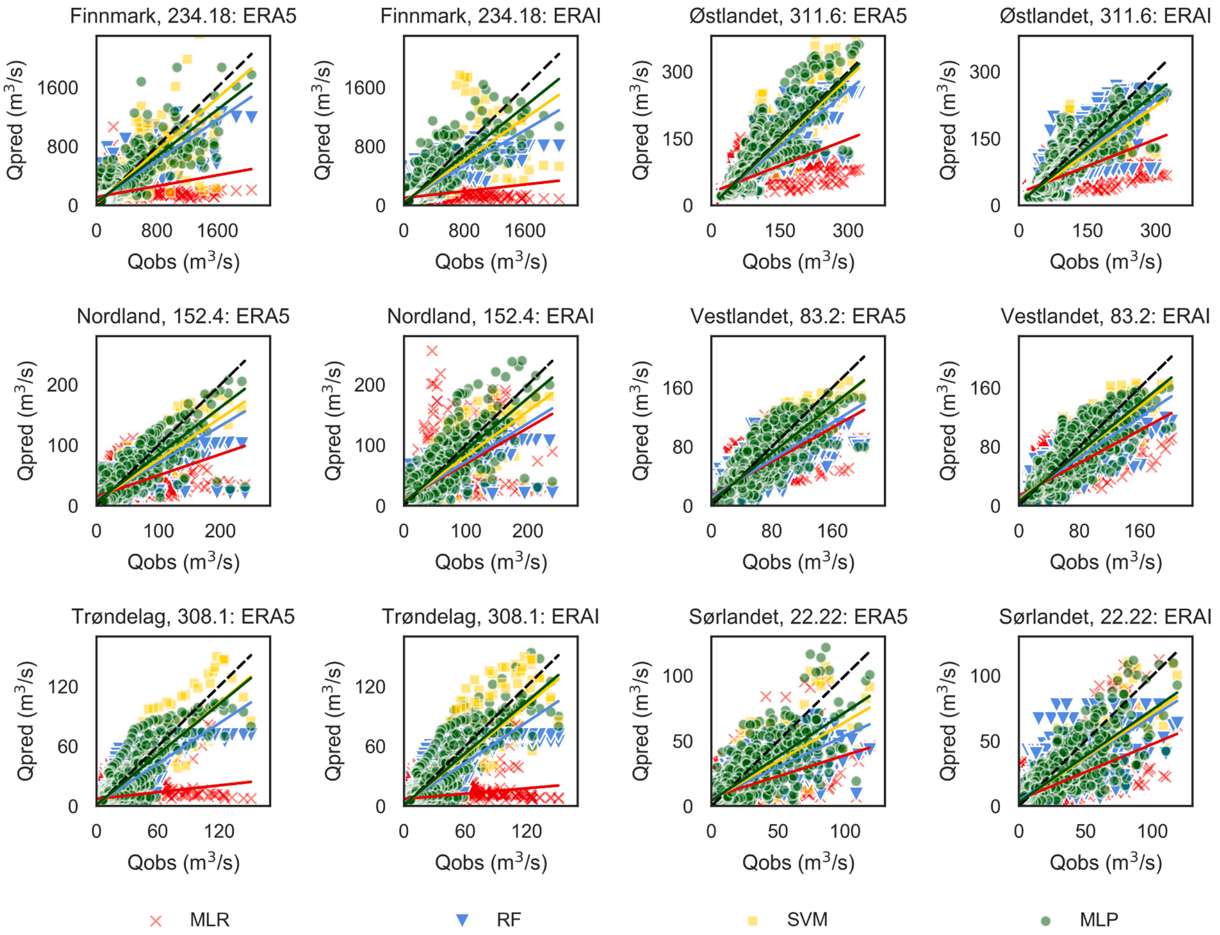


Fig. 11. Scatter plots of observed versus reconstructed flows by MLR (red crosses), RF (blue triangles), SVM (yellow squares) and MLP (green circles) using identified drivers . The black dashed line is the 45° line corresponding to a perfect model. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

moving window; this number could be considered a function of catchment size, dominant streamflow-generating mechanism and annual flow, giving larger candidate retrieval per moving window in smaller, rainfall-driven, fast-responding catchments.

It should be emphasized that the method directly linking large-scale atmospheric variables to streamflow on daily level presented in this paper is not intended as a replacement for traditional hydrological modelling tools. This study has identified major large-scale atmospheric drivers using traditional machine learning techniques. For streamflow prediction, the importance of catchment characteristics and their role in transforming atmospheric inputs into a spatial and temporal pattern of streamflow cannot be overstated. However, the possibility of establishing such direct links between the atmosphere and the hydrological regimes on a daily level unlocks opportunities for direct downscaling to daily streamflow from climate model outputs. This will provide a deeper understanding of how climatic changes in large-scale atmospheric circulation can drive changes in streamflow at the catchment scale. As such, machine learning techniques can supplement conventional hydrological models by providing a complementary approach to interpreting potential future changes – directly from global circulation models. Thus, a complementary approach that shortens the state-of the-art modelling chain – with downscaling, bias-correction and hydrological modelling – will allow for evaluation of a larger ensemble of models and hence provide better uncertainty estimates, as well as avoid some of the potential errors and biases introduced in the modelling chain. As a result, a stronger physical basis for explaining patterns of non-stationarity in the current and past climate may be obtained and thus further be used to distinguish natural variability from climate change-induced trends.

4.3. *Potential and limitations*

Snowmelt-driven floods in Norway are decreasing in magnitude and frequency (Vormoor et al., 2016). This trend is expected to continue into the future, as rising temperatures lead to less snow accumulation (Hanssen-Bauer et al., 2017; Lawrence, 2020), also causing a shift in the

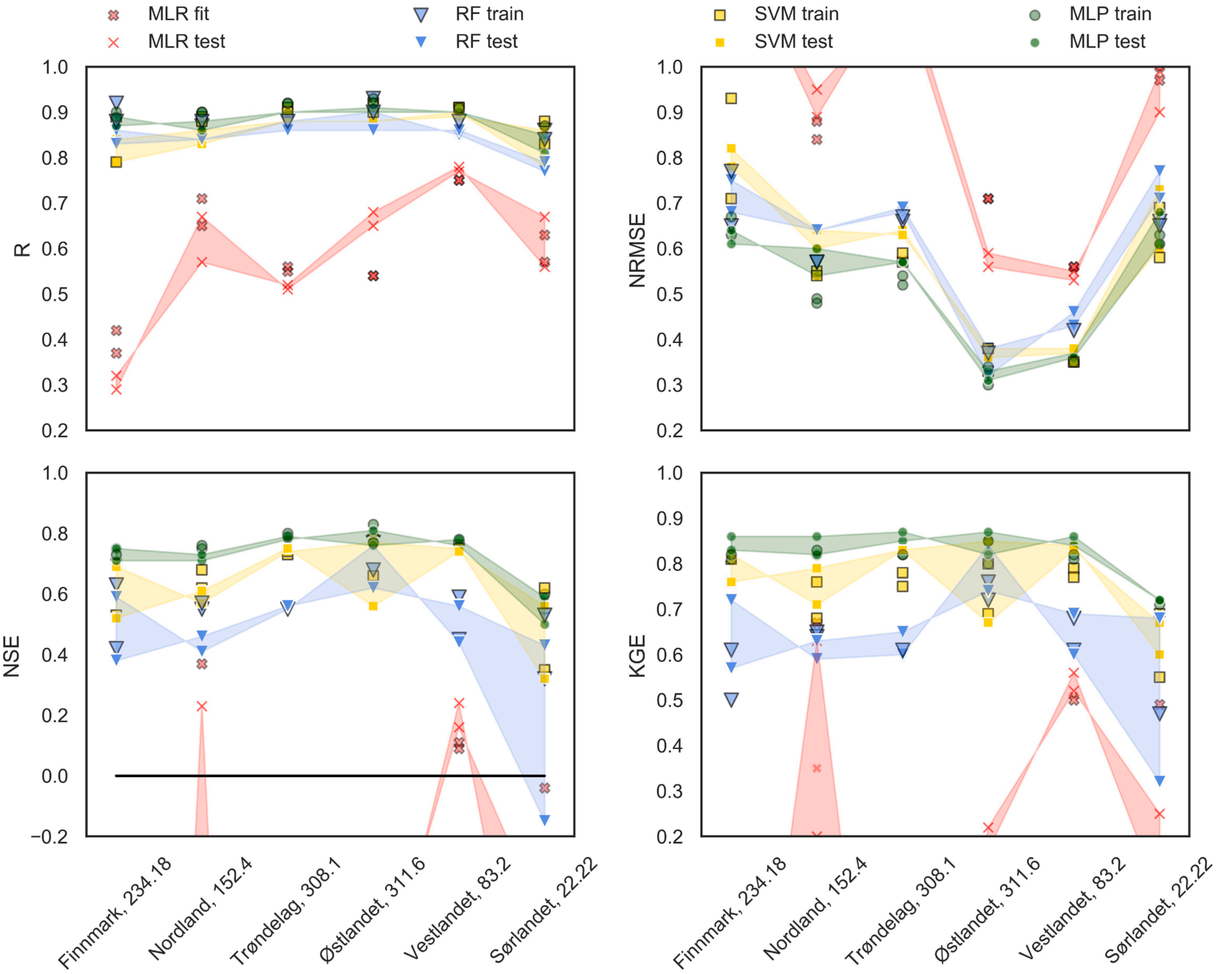


Fig. 12. Range of scores obtained on the four metrics, Pearson's r (R), normalized root-mean-square error (NRMSE), Nash-Sutcliffe Efficiency (NSE) and Kling-Gupta Efficiency (KGE) on the testing set. The spread gives a measure of the consistency between scores obtained with ERA5 and ERAI by MLR (red crosses), RF (blue triangles), SVM (yellow squares) and MLP (green circles) on the testing set. The scores obtained during training are displayed with black marker outlines. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

timing of spring floods. As snowmelt-driven floods decrease, a transient change towards predominantly rainfall-driven streamflow generation also in northern or mountainous areas can be expected. However, this transition need not be linear or monotonic due to the competing effects of increases in winter precipitation versus temperature. While none of the hydrological stations used in this study showed statistically significant trends, several hydrological stations in Europe have shown both statistically significant increases and decreases (Blöschl et al., 2019); future applications of the methodology presented in this paper may therefore need to include learning of trends as a minimum requirement to cope with non-stationarity.

In the current model setup, only one atmospheric column per station was considered representative of the relevant large-scale atmospheric circulation. A more elaborate approach may involve a grid search with the use of *meta*-heuristics, resulting in more than one representative atmospheric column per station and hence a more spatially distributed identification of major drivers. This may be particularly useful in large catchments with contrasting characteristics, in which daily streamflow results from several distinct and independent drivers that may form compound events. An example of this would be partly mountainous catchments with low-lying valleys where rivers directly connect to the sea so that drainage is affected by pressure. While climate indices may – complementary to large-scale atmospheric variables – feed as model input to identify major drivers, care should be taken, as robust representation of large-scale atmospheric circulation is needed also in a changing climate.

Despite the fact that existing discipline knowledge may be used to assess the consistency and interpretability of feature selections as employed in this study, a clear separation of correlation and causation cannot be inferred from the model structures. This becomes increasingly problematic as the model complexity increases. Furthermore, ensuring that the maximum peaks were included in the training data facilitated a fair comparison of the generalization obtained by the three machine learning models. In other words, extrapolation techniques in each algorithm were not explicitly addressed in this study, but should be given attention in future studies.

The findings of this paper demonstrate the potential for direct downscaling from large-scale atmospheric variables to daily streamflow with shallow neural networks. This calls for an investigation of gain in accuracy with deep neural networks. Despite a growing interest in the scientific community, deep learning is largely unexplored in the field of hydrology (Ardabili et al., 2020), and the number of available algorithms and model structures are increasing – particularly through hybridization. As transparency tends to decrease reciprocally with model complexity, a robust collection of benchmark studies with traditional machine learning techniques is needed. With a sample of six hydrological stations from catchments with low anthropogenic influence, the generalization of the findings of this paper is limited. This calls for further benchmark studies involving more hydrological stations; for the case of Norway, the additional 34 regionally filtered stations from the Norwegian Hydrological Reference Dataset for Climate Impact Studies (Fleig, 2013) may be tested with the methodology presented here. As such, this paper provides a step towards a collection of robust benchmarks for deep learning models in the context of daily streamflow reconstruction and statistical downscaling from large-scale atmospheric variables in Norway.

5. Conclusion

This paper identified major drivers of daily streamflow in snowmelt-driven and rainfall-driven flood regimes in Norway from large-scale atmospheric variables in two reanalysis datasets of fine (0.25°) and coarse (0.75°) spatial resolution. Consistency between the feature selections from the two reanalysis datasets was confirmed. The translation from large-scale atmospheric forcing to daily streamflow using the identified drivers was investigated in relation to model complexity with three machine learning models: random forest (RF), support vector machine (SVM) for regression and multilayer perceptron (MLP). A sampled screening of six hydrological stations located in catchments with a low degree of anthropogenic influence was presented. This paper provides a first step towards applying machine learning for direct downscaling from large-scale atmospheric circulation on a daily temporal scale in Norway and may hence serve as a benchmark for later developments of deep learning models.

In answering the research question posed introductorily, the following is concluded:

Major drivers of daily streamflow in each of the six investigated catchments reflect unique forcing-response relationships from the atmosphere, but also exhibit similarities within dominantly snowmelt-driven and rainfall-driven flood regimes respectively. In catchments where high flows to a large extent are generated by snowmelt (234.18, Finnmark; 308.1, Trøndelag; and 311.6, Østlandet), two distinct aggregation periods of boundary layer and near-surface temperature variates were evident at one to three weeks and two to three months, respectively representing melting and seasonality. A consistent aggregation window in between these two dominant aggregation periods was also found, reflecting snow accumulation with moisture and temperature variates. In catchments where high flows primarily are generated by rainfall (152.4, Nordland; 83.2, Vestlandet; and 22.22, Sørlandet), the dominant aggregation period does not extend beyond 3 weeks. In these catchments, the identified major drivers are moisture, wind and temperature variates in the boundary layer, with less short-term influence from variables in the middle and upper troposphere. Seasonality is not represented through temperature variates, but rather moisture and wind variates with aggregation windows around three months.

Increasing accuracy with model complexity was found for all the six investigated catchments. Furthermore, the most complex model, MLP, was found to be the most robust machine learning model with the lowest drop in Nash-Sutcliffe Efficiency (NSE) from training to testing. In all but the southernmost, and smallest, catchment, MLP obtained an NSE ranging from 0.71 to 0.81, with an average drop of less than 0.02. This supports previous findings stating that neural networks handle non-linearity in complex systems better than linear or piecewise linear approaches. The inter-model and intra-model differences in performance indicate the strength of the relationship between direct forcing from large-scale atmospheric circulation and daily streamflow and may hence be used to filter out suitable catchments for direct downscaling in the context of hydrological climate-impact modelling. However, since streamflow is the product of both atmospheric forcing and catchment characteristics, the latter must be explicitly included in streamflow prediction using the identified major drivers. Among the six investigated catchments, station 311.6 in Østlandet was determined most suitable for direct downscaling, while station 22.22 in Sørlandet was determined unsuitable without further investigation of the relative importance of catchment characteristics, sub-grid features and local variability by traditional hydrological modelling approaches.

Future research should focus on further development of benchmark studies and move towards exploration of deep learning for daily streamflow reconstruction with major drivers identified from large-scale atmospheric forcing and catchment characteristics. Increasing the number of hydrological stations representing the various snowmelt-driven and rainfall-driven flood regimes will particularly aid in assessing the scalability of the findings presented in this paper. Furthermore, clustering of hydrological stations based on catchment characteristics may allow the development of a transient model structure, in which [changes in] dominant streamflow-generating mechanisms or timing of annual peaks may feed into the automated feature selection procedure. As such, the input variable selections may be directed towards features and aggregation windows characteristic of rainfall-driven flood regimes in line with decreasing snowmelt-driven floods. For all of these potential future studies, the work presented here may serve as a benchmark.

CRediT authorship contribution statement

Jenny Sjåstad Hagen: Conceptualization, Methodology, Software, Validation, Formal analysis, Data curation, Writing - original draft, Writing - review & editing, Visualization. **Etienne Leblois:** Conceptualization, Methodology, Investigation, Data curation. **Deborah Lawrence:** Conceptualization, Writing - original draft, Writing - review & editing, Supervision. **Dimitri Solomatine:** Conceptualization, Supervision. **Asgeir Sorteberg:** Conceptualization, Methodology, Resources, Funding acquisition, Supervision.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

The authors would like to thank the Norwegian Water Resources and Energy Directorate (NVE), The European Centre for Medium Range Weather Forecasts (ECMWF) and the Climate Data Store (CDS) for providing access to data used in this study. The authors would also like to thank Marvin Kähnert for proofreading and useful comments. This research was supported by the Geophysical Institute, University of Bergen, and the Bjerknes Centre for Climate Research through the Climate Hazards and Extremes (CHEX) project and is also part of the NFR KeyClim project (no. 295046). The authors developed this collaboration following discussions at a conference sponsored by the Norwegian Hydrological Council.

Appendix A

Table A1

Table A1
List of symbols for atmospheric and surface variables from reanalysis data.

Symbol	Variable name
Cloud	Cloud cover
td2m	Dew point temperature at 2 m above surface
Heat	Heating
LHeat	Latent heat
LW	Outgoing longwave radiation
Precip	Precipitation
RH	Relative humidity
SH	Specific humidity
SHeat	Sensible heat
SLP	Mean sea level pressure
SST	Sea surface temperature
SW	Incoming shortwave radiation
T	Temperature
t2m	Air temperature 2 m above surface
Tcv	Total column vapor
Tcw	Total column water
U	Eastward wind
V	Northward wind
W	Vertical wind
Z	Geopotential

References

Abbott M.B. Hydroinformatics: Information Technology and the Aquatic Environment - Michael B 1991 Abbott - Google Books Avebury Technical..

Abrahart, R.J., Anctil, F., Coulibaly, P., Dawson, C.W., Mount, N.J., See, L.M., Shamseldin, A.Y., Solomatine, D.P., Toth, E., Wilby, R.L., 2012. Two decades of anarchy? Emerging themes and outstanding challenges for neural network river forecasting. Prog. Phys. Geogr. 36, 480–513. https://doi.org/10.1177/0309133312444943.

Adnan, R.M., Liang, Z., Trajkovic, S., Zounemat-Kermani, M., Li, B., Kisi, O., 2019. Daily streamflow prediction using optimally pruned extreme learning machine. J. Hydrol. 577 https://doi.org/10.1016/j.jhydrol.2019.123981.

S. Ardabili A. Mosavi M. Dehghani A.R. Várkonyi-Kóczy Deep Learning and Machine Learning in Hydrological Processes Climate Change and Earth Systems a Systematic Review 2020 Springer 52 62 10.1007/978-3-030-36841-8_5.

G. Blöschl J. Hall A. Viglione R.A.P. Perdigão J. Parajka B. Merz D. Lun B. Arheimer G.T. Aronica A. Bilibashi M. Boháč O. Bonacci M. Borga I. Čanjevac A. Castellarin G.B. Chirico P. Claps N. Frolova D. Ganora L. Gorbachova A. Gül J. Hannaford S. Harrigan M. Kireeva A. Kiss T.R. Kjeldsen S. Kohnová J.J. Koskela O. Ledvinka N. Macdonald M. Mavrova-Guirguinova L. Mediero R. Merz P. Molnar A. Montanari C. Murphy M. Osuch V. Ovcharuk I. Radevski J.L. Salinas E. Sauquet M. Šraj J. Szolgay E. Volpi D. Wilson K. Zaimi N. Živković Changing climate both increases and decreases European river floods Nature 2019 10.1038/s41586-019-1495-6.

Cannon, A.J., Whitfield, P.H., 2002. Downscaling recent streamflow conditions in British Columbia, Canada using ensemble neural network models. J. Hydrol. 259, 136–151. https://doi.org/10.1016/S0022-1694(01)00581-9.

Chu, H., Wei, J., Wu, W., 2020. Streamflow prediction using LASSO-FCM-DBN approach based on hydro-meteorological condition classification. J. Hydrol. 580 https://doi.org/10.1016/j.jhydrol.2019.124253.

Cortes, C., Vapnik, V., 1995. Support-vector networks. Mach. Learn. 20, 273–297. https://doi.org/10.1007/BF00994018.

Das, J., Nanduri, U.V., 2018. Assessment and evaluation of potential climate change impact on monsoon flows using machine learning technique over Wainganga River basin. India. Hydrol. Sci. J. 63, 1020–1046. https://doi.org/10.1080/02626667.2018.1469757.

D.P. Dee S.M. Uppala A.J. Simmons P. Berrisford P. Poli S. Kobayashi U. Andrae M.A. Balmaseda G. Balsamo P. Bauer P. Bechtold A.C.M. Beljaars L. van de Berg J. Bidlot N. Bormann C. Delsol R. Dragani M. Fuentes A.J. Geer L. Haimberger S.B. Healy H. Hersbach E.V. Hólm L. Isaksen P. Kållberg M. Köhler M. Matricardi A.P. Mcnally B. M. Monge-Sanz J.J. Morcrette B.K. Park C. Peubey P. de Rosnay C. Tavolato J.N. Thépaut F. Vitart The ERA-Interim reanalysis: Configuration and performance of the data assimilation system Q. J. R. Meteorol. Soc. 137 2011 553 597 10.1002/qj.828.

Engeland K. Ed., Schlichting, L., Randen, F., Nordtun, K.S., Reitan, T., Wang, T., Holmqvist, E., Voksø, A., Eide, V 2016 Flomdata Utvalg og kvalitetssikring av flomdata for flomfrekvensanalyser Oslo.

Fleig Norwegian Hydrological Reference Dataset for Climate Change Studies 2013 Oslo.

Ghosh, S., Mujumdar, P.P., 2008. Statistical downscaling of GCM simulations to streamflow using relevance vector machine. Adv. Water Resour. 31, 132–146. https://doi.org/10.1016/j.advwatres.2007.07.005.

Gupta, H.V., Kling, H., Yilmaz, K.K., Martinez, G.F., 2009. Decomposition of the mean squared error and NSE performance criteria: Implications for improving hydrological modelling. J. Hydrol. 377, 80–91. https://doi.org/10.1016/j.jhydrol.2009.08.003.

Hanssen-Bauer, I., Førland, E.J., Haddeland, I., Hisdal, H., Mayer, S., Nesje, A., Nilsen, J. E.Ø., Sandven, S., Sandø, A.B., Sorteberg, A., Ådlandsvik, B., 2017. Climate in Norway 2100 - a knowledge base for climate adaptation.

H. Hersbach B. Bell P. Berrisford S. Hirahara A. Horányi J. Muñoz-Sabater J. Nicolas C. Peubey R. Radu D. Schepers A. Simmons C. Soci S. Abdalla X. Abellan G. Balsamo P. Bechtold G. Biavati J. Bidlot M. Bonavita G. De Chiara P. Dahlgren D. Dee M. Diamantakis R. Dragani J. Flemming R. Forbes M. Fuentes A. Geer L. Haimberger S. Healy R.J. Hogan E. Hólm M. Janisková S. Keeley P. Laloyaux P. Lopez C. Lupu G. Radnoti P. de Rosnay I. Rozum F. Vamborg S. Villaume J.N. Thépaut The ERA5 global reanalysis 2020 R. Meteorol. Soc Q. J 10.1002/qj.3803.

Huang, S., Lawrence, D., Irene Brox, N., Li, H., 2020. Direct statistical downscaling of monthly streamflow from atmospheric variables in catchments with differing contributions from snowmelt. Int. J. Climatol. joc.6878 https://doi.org/10.1002/joc.6878.

M. Hussain I. Mahmud pyMannKendall: a python package for non parametric Mann Kendall family of trend tests J. Open Source Softw. 4 2019 1556 https://doi.org/10.21105/joss.01556.

Jiang, S., Zheng, Y., Solomatine, D., 2020. Improving AI system awareness of geoscience knowledge: symbiotic integration of physical approaches and deep learning. Geophys. Res. Lett. 47 https://doi.org/10.1029/2020GL088229.

Kingston, D.G., Lawler, D.M., McGregor, G.R., 2006. Linkages between atmospheric circulation, climate and streamflow in the northern North Atlantic: research prospects. Prog. Phys. Geogr. Earth Environ. 30, 143–174. https://doi.org/10.1191/0309133306 pp471ra.

Knoben, W.J.M., Freer, J.E., Woods, R.A., 2019. Technical note: Inherent benchmark or not? Comparing Nash-Sutcliffe and Kling-Gupta efficiency scores. Hydrol. Earth Syst. Sci. 23, 4323–4331. https://doi.org/10.5194/hess-23-4323-2019.

Kottek, M., Grieser, J., Beck, C., Rudolf, B., Rubel, F., 2006. World Map of the Köppen-Geiger climate classification updated. Meteorol. Zeitschrift 259–263. https://doi.org/10.1127/0941-2948/2006/0130.

Kratzert, F., Klotz, D., Brenner, C., Schulz, K., Herrnegger, M., 2018. Rainfall–runoff modelling using Long Short-Term Memory (LSTM) networks. Hydrol. Earth Syst. Sci. 22, 6005–6022. https://doi.org/10.5194/hess-22-6005-2018.

Kratzert, F., Klotz, D., Herrnegger, M., Sampson, A.K., Hochreiter, S., Nearing, G.S., 2019. toward improved predictions in ungauged basins: exploiting the power of machine learning. Water Resour. Res. 55, 11344–11354. https://doi.org/10.1029/2019WR026065.

Lawrence, D., 2020. Uncertainty introduced by flood frequency analysis in projections for changes in flood magnitudes under a future climate in Norway. J. Hydrol. Reg. Stud. 28 https://doi.org/10.1016/j.ejrh.2020.100675.

Liao, S., Liu, Z., Liu, B., Cheng, C., Jin, X., Zhao, Z., 2019. Multi-step ahead daily inflow forecasting using ERA-Interim reanalysis dataset based on gradient boosting regression trees. Hydrol. Earth Syst. Sci. Discuss. 1–28 https://doi.org/10.5194/hess-2019-610.

Moradi, A.M., Dariane, A.B., Yang, G., Block, P., 2020. Long-range reservoir inflow forecasts using large-scale climate predictors. Int. J. Climatol. joc.6526 https://doi.org/10.1002/joc.6526.

Mosavi, A., Ozturk, P., Chau, K., Mosavi, A., Ozturk, P., Chau, K., 2018. Flood prediction using machine learning models: literature review. Water 10, 1536. https://doi.org/10.3390/w10111536.

Nash, J.E., Sutcliffe, J.V., 1970. River flow forecasting through conceptual models part I — a discussion of principles. J. Hydrol. 10, 282–290. https://doi.org/10.1016/0022-1694(70)90255-6.

Nilsson, P., Uvo, C.B., Landman, W.A., Nguyen, T.D., 2008. Downscaling of GCM forecasts to streamflow over Scandinavia. Hydrol. Res. 39, 17–26. https://doi.org/10.2166/nh.2008.027.

Nogueira, M., 2020. Inter-comparison of ERA-5, ERA-interim and GPCP rainfall over the last 40 years: process-based analysis of systematic and random differences. J. Hydrol. 583 https://doi.org/10.1016/j.jhydrol.2020.124632.

Norwegian Ministry of Petroleum and Energy Electricity production - Energifakta Norge [WWW Document] https://energifaktanorge.no/en/norsk-energiforsyning/kraftproduksjon/ 2015 accessed 8.25.20.

Okkan, U., Inan, G., 2015. Statistical downscaling of monthly reservoir inflows for Kemer watershed in Turkey: use of machine learning methods, multiple GCMs and emission scenarios. Int. J. Climatol. 35, 3274–3295. https://doi.org/10.1002/joc.4206.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., 2011. Scikit-learn: machine learning in Python. J. Mach. Learn. Res. 12, 2825–2830.

Rasouli, K., Hsieh, W.W., Cannon, A.J., 2012. Daily streamflow forecasting by machine learning methods with weather and climate inputs. J. Hydrol. 414–415, 284–293. https://doi.org/10.1016/j.jhydrol.2011.10.039.

Reichstein, M., Camps-Valls, G., Stevens, B., Jung, M., Denzler, J., Carvalhais, N., Prabhat, 2019. Deep learning and process understanding for data-driven Earth system science. Nature 566, 195–204. https://doi.org/10.1038/s41586-019-0912-1.

Ren, W., Yang, T., Shi, P., Xu, C., yu, Zhang, K., Zhou, X., Shao, Q., Ciais, P, 2018. A probabilistic method for streamflow projection and associated uncertainty analysis in a data sparse alpine region. Glob. Planet. Change 165, 100–113. https://doi.org/10.1016/j.gloplacha.2018.03.011.

Sachindra, D.A., Huang, F., Barton, A., Perera, B.J.C., 2013. Least square support vector and multi-linear regression for statistically downscaling general circulation model outputs to catchment streamflows. Int. J. Climatol. 33, 1087–1106. https://doi.org/10.1002/joc.3493.

Sahoo, A., Sen, D., 2017. Assessing climate changes in california, using support vector machine in statistical downscaling. IUP J. Comput. Sci. 11, 7–25.

Sivakumar, B., Berndtsson, R., Olsson, J., Jinno, K., 2001. Evidence of chaos in the rainfall-runoff process. Hydrol. Sci. J. 46, 131–145. https://doi.org/10.1080/02626660109492805.

Thapa, S., Zhao, Z., Li, B., Lu, L., Fu, D., Shi, X., Tang, B., Qi, H., 2020. Snowmelt-driven streamflow prediction using machine learning techniques (LSTM, NARX, GPR, and SVR). Water 12, 1734. https://doi.org/10.3390/w12061734.

Tisseuil, C., Vrac, M., Lek, S., Wade, A.J., 2010. Statistical downscaling of river flows. J. Hydrol. 385, 279–291. https://doi.org/10.1016/j.jhydrol.2010.02.030.

Tongal, H., Booij, M.J., 2018. Simulation and forecasting of streamflows using machine learning models coupled with base flow separation. J. Hydrol. 564, 266–282. https://doi.org/10.1016/j.jhydrol.2018.07.004.

Vojinovic, Z., Abbott, M.B., 2017. Twenty-five years of hydroinformatics. Water (Switzerland). https://doi.org/10.3390/w9010059.

K. Vormoor D. Lawrence M. Heistermann A. Bronstert Climate change impacts on the seasonality and generation processes of floods - projections and uncertainties for catchments with mixed snowmelt/rainfall regimes Hydrol. Earth Syst. Sci. 19 2015 913 931 10.5194/hess-19-913-2015.

Vormoor, K., Lawrence, D., Schlichting, L., Wilson, D., Wong, W.K., 2016. Evidence for changes in the magnitude and frequency of observed rainfall vs. snowmelt driven floods in Norway. J. Hydrol. 538, 33–48. https://doi.org/10.1016/J.JHYDROL.2016.03.066.

Xiang, Z., Yan, J., Demir, I., 2020. A rainfall-runoff model with LSTM-based sequence-to-sequence learning. Water Resour. Res. 56 https://doi.org/10.1029/2019WR025326.

Yaseen, Z.M., El-shafie, A., Jaafar, O., Afan, H.A., Sayl, K.N., 2015. Artificial intelligence based models for stream-flow forecasting: 2000–2015. J. Hydrol. 530, 829–844. https://doi.org/10.1016/J.JHYDROL.2015.10.038.