Regional median flood estimation with generalized additive models: model selection across durations

Danielle M. Barna[*,1,3], Kolbjørn Engeland[1], Thomas Kneib[5], Thordis L. Thorarinsdottir[2,4], and Chong-Yu Xu[3]

[1]Norwegian Water Resources and Energy Directorate, P.O. Box 5091 Majorstua, NO-0301 Oslo, Norway
[2]Norwegian Computing Centre, Oslo, Norway
[3]Department of Geosciences, University of Oslo, Oslo, Norway.
[4]Department of Mathematics, University of Oslo, Oslo, Norway.
[5]Chair of Statistics, University of Göttingen, Göttingen, Germany.

Correspondence: Danielle M. Barna (daba@nve.no)

ABSTRACT

Estimating flood quantiles in ungauged basins often relies on regression methods. When flood retention is crucial, such as in floodplain management and reservoir design, flood quantile estimates are needed for multiple durations. This presents a challenge for regression models as the relationship between catchment descriptors and response may vary across durations. Generalized Additive Models (GAMs) are suitable for this situation, offering flexible, semi-parametric modeling and visualization of predictor-response relationships. However, selecting predictors for GAMs can be difficult due to the nature of catchment descriptor datasets. To address this, we use a machine-learning algorithm to flag promising predictors for expert assessment. In this study, we develop a GAM for median flood estimation, focusing on duration-specific differences in how catchment descriptors affect the median flood. We assess the GAM's adequacy through predictive performance and reliability. We compare the GAM to two benchmarks: Norway's existing log-linear model for median flood estimation and an extreme gradient boosting tree ensemble (XGBoost). The GAM's predictive accuracy and reliability match or exceed those of the benchmarks for all studied durations. We find duration-specific differences in the relationship between median flood and descriptors like effective lake percentage and catchment shape, with statistically significant performance declines if these differences are ignored.

KEYWORDS

flood quantiles at multiple durations, generalized additive models, index flood estimation, regional flood frequency analysis, regression, XGBoost

HIGHLIGHTS

- Flexible modeling: GAMs auto-adapt regression relationships to fit different durations.
- Predictor selection: machine learning flags promising predictors for use in GAMs.
- Comparative performance: the GAM matches or exceeds the benchmark models on both predictive accuracy and reliability at all durations.
- Impact of ignoring duration-specific differences: reduced predictive performance when models developed for longer durations are directly applied to shorter durations.

INTRODUCTION

Estimating flood quantiles in ungauged catchments is a frequent challenge in hydrology and is crucial to many tasks related to design of infrastructure, emergency and land use planning. A common method for flood quantile estimation at ungauged catchments is regional flood frequency analysis (RFFA), which uses catchment characteristics–that is, descriptors of the physical properties of the catchment such as, for example, area, lake percentage, average elevation and total river length–as well as climate characteristics like precipitation and temperature to infer flood quantiles at a specified ungauged location. Often this inference is based on the concept that spatial variations in flood statistics are closely linked with regional catchment and climate characteristics and is achieved through regression based methods (Robson and Reed, 1999).

This study focuses on constructing regression models for the median annual maximum (index) flood at multiple flood durations. Many hydrologic applications where flood retention is important, e.g. floodplain management and reservoir design, require flood quantile estimates at different durations (Lamontagne et al., 2012). For example, hourly durations are relevant for smaller hydropower stations and stream inflow to smaller reservoirs, while longer durations (that is, durations ranging from 24 hours to a few days) are relevant for large reservoirs. We consider a set of flood durations ranging from 1 hour to 48 hours in this study (1 hour, 6 hour, 12 hour, 18 hour, 24 hour, 36 hour and 48 hour durations).

Design values that involve duration can be treated under an event-based, multivariate framework that focuses on explicitly modeling the dependency structure between different multivariate aspects of events (i.e. "treating the duration as a variable"; see, for example, Gräler et al. (2013) and Requena et al. (2016)). This is a useful approach, in particular when multivariate design events are desired. However, engineering design for retention-specific applications often requires flood volumes for pre-determined durations, sometimes averaged over different flood events, rather than the multivariate variability of specific flood events. This requires an aggregation-based approach to duration. We take an aggregation-based approach to duration in this study. Durations therefore represent the total flow volume observed over a desired time window.

Using regression models to estimate flood quantiles at multiple durations requires special considerations. Regression models used for median flood estimation are typically *parametric* regression models (e.g. linear, log-linear, nonlinear, or generalized linear models), meaning the models rely on a parametric description of what is called the functional form between predictors

and response. Parametric models are easy to interpret and estimate and are therefore widely used. However, in the particular case where we predict the median flood, using a parametric model developed for one duration on a different duration assumes that, although magnitude and direction may change when regression coefficients are re-estimated or scaled for different durations, the functional form of the relationship between catchment descriptors and the median flood remains the same across durations. This has practical implications: if the relationship between catchment descriptors and the median flood changes based on how fast the volume of water arrives, a parametric regression model developed for one duration may provide suboptimal results when applied to another duration.

Overall, there is a gap in regional flood frequency analysis when it comes to assessing regression models for flood quantiles at multiple durations. In this study, we build models independently for each duration and examine whether there are statistically significant differences in predictive accuracy, reliability, and fitted relationships across different durations and different types of regression models. This requires development of interpretable, non- or semi-parametric (data-driven) regression models for median flood prediction at different durations. A requirement is that the new models must be able to be compared to the existing parametric log-linear model for median flood estimation in Norwegian catchments on a variety of metrics and assessments relevant for practical application. This requires accessible uncertainty estimates and access to the full predictive distribution. Thus, although full machine learning models are a powerful and flexible option for data-driven model architectures, they are not appropriate in this situation. Instead, we implement a particular type of regression model called a generalized additive model (GAM).

GAMs are semi-parametric extensions of the linear regression model that can account for complex nonlinear relationships between predictors and response. The semi-parametric nature of GAMs means they are highly flexible: the data determine the form of the relationship between the response and the predictors rather than assuming some form of parametric relationship, e.g. transformation of predictors. The main use case of GAMs lies in applications where a nonlinear relationship between the predictor and response needs to be defined or established; however, if the relationship is linear, the smooth function that defines the relationship between predictor and response will recover the linear relationship. This offers a potential simplification of the modeling process as we do not need to identify appropriate predictor transformations. Furthermore, the fact we no longer have to specify a functional form of the predictor-response relationship has potential to generate relationships that better represent the underlying data. This allows for statistical analyses that focus on identification and description of the data-driven relationships between predictor and response. The relationships identified can in some cases increase our understanding of hydrologic systems, although the reality of the functional relationships should always be established by theory or process-based models outside of the statistical analysis.

Specific application of GAMs for estimation of flood quantiles is generally first attributed to Chebana et al. (2014). Here GAMs were used to estimate the quantiles corresponding to the 10, 50 and 100 year return periods and the models were compared to log-linear models on a variety of different regional groupings. The GAMs were found to have improved predictive performance and the flexibility of the GAMs reduced the need to split into hydrologically homogenous regions. Other examples of GAMs used for flood quantile regression are the comparative studies of Msilini et al. (2022) and Rahman et al. (2018), which compared GAMs to traditional log-linear regression approaches, among others. Rahman et al. (2018), in line with Chebana

et al. (2014), found that GAMs typically outperformed log-linear models, even without constraining the GAM to hydrologically similar neighborhoods or regions of influence. In an application that looked at quantile-specific performance differences, Noor et al. (2022) compared a GAM to a linear quantile regression technique and found the GAM resulted in improved performance, but only on quantiles associated with small return periods of up to ten years.

As with any data-driven model, careful predictor selection is necessary for GAMs to prevent overfitting and ensure robust predictive performance. The mathematical structure of GAMs allows for very simple but effective *shrinkage-based* predictor selection methods to be used; see Marra and Wood (2011) for development of the methodology and performance comparison to alternative methods (e.g. stepwise selection). Shrinkage-based methods offer the benefits of subset selection while allowing predictor selection to be accomplished in a single step, and are particularly appealing in a practical context as they allow for predictor selection uncertainty to be included in the final model.

However, the usefulness of shrinkage methods is limited to relatively small sets of uncorrelated predictors. This is problematic as regional hydrologic datasets typically contain many correlated predictors. For this reason, current applications either solely use backwards stepwise selection (which cannot provide estimates of predictor selection uncertainty) (Chebana et al., 2014; Rahman et al., 2018; Noor et al., 2022; Msilini et al., 2022) or use expert knowledge to pre-select a small, uncorrelated set of predictors that can then be validated using shrinkage methods (Dubos et al., 2022).

In this study, we take the second approach, and use in addition an existing machine learning-based algorithm to identify promising predictors for expert assessment. We chose the tree-based Iterative Input Selection algorithm (IIS) presented in Galelli and Castelletti (2013) and applied in, for example, Prasad et al. (2017), He et al. (2022) and Pesantez et al. (2020). Although this preliminary step of using machine learning to flag promising predictors is complementary to, and not necessary for, our analysis, we outline the setup of the algorithm anyway. A machine-learning based algorithm is in no way a "silver bullet" that is guaranteed to identify predictors that will be useful to a GAM. The setup required a series of careful choices that complement GAM development. We therefore find it useful to report the steps taken to potentially assist with predictor selection in GAMs, which is in general challenging and a major roadblock to setting up analyses such as this one.

The primary objective of our study is detection and description of the functional relationships between the median flood and catchment covariates at different durations. Here we assume that, while the relationship between the covariates and the median flood may vary with duration, the covariates themselves remain constant across different durations. The accuracy of this explainable approach is dependent on the fitted GAM being adequate. Thus the secondary objective of our study is prediction of the median flood at ungauged locations, where predictive performance and reliability at ungauged locations are used as proxies for adequacy of the GAM. We use two benchmark models to establish predictive performance. These are the existing log-linear model for median flood estimation in Norway and a gradient-boosted tree ensemble (XGBoost). XGBoost has established use in hydrology (Zounemat-Kermani et al., 2021) and is applied in, for example, Laimighofer et al. (2022a) and Ni et al. (2020). As part of what distinguishes the GAM from the log-linear model is the flexible, data-driven nature of the response relationship, it is useful to have a comparison point from a fully data-driven model such as XGBoost.

The following research questions will be addressed:

1. Can the GAM achieve comparable or improved performance compared to the benchmark models across different durations?

2. Can we identify and describe duration-specific differences in how catchment covariates influence the median flood? How impactful are these differences? (i.e. if we ignore them, what is the impact on predictive performance?).

Our analysis will be performed on annual maximum data since flood guidelines in Norway pertain to annual maximum values.

The remainder of the paper is organized as follows: we first introduce the flood data and catchment descriptors. The following section presents an outline of the study design. The methods section presents the GAM used in this study and summarizes the chosen predictor selection approach. This section also summarizes the two reference models and the evaluation methods used to assess all models in the study. The results section presents the predictive performance and model reliability results as well as the functional relationships identified by the GAM. The paper finishes with a discussion and conclusions.

DATA

Flood data from 232 gauging stations across Norway were used in this study (Fig. 1). The stations exhibit a diversity of hydroclimatic regimes relative to Nordic catchments. The spatial distribution of temperature and precipitation regimes in Nordic countries is primarily influenced by climatological gradients associated with latitude, topography, and proximity to the coastal zone; the diverse topography and wide range of latitudes in Norway make it a suitable location for regionalization studies in the Nordic region.

In Norway the two major flood generating processes are snowmelt and rainfall. The regional importance of snowmelt as a runoff generating process varies greatly due to differences in the temperature regime, snowpack volumes and the snow season across the country. Inland and northern regions are those primarily driven by snowmelt and experience prominent high flows during spring and summer, while western and coastal regions are primarily driven by rainfall and experience high flows during autumn and winter. However, local climate and mixed or transitional flood regimes mean these regional patterns exhibit great variability, and seasonal patterns are not very distinct in rainfall-driven catchments (Vormoor et al., 2016).

The observed streamflow time series were obtained from the national hydrological database Hydra II hosted by the Norwegian Water Resources and Energy Directorate (NVE). The streamflow records have at least 20 years of quality controlled data for periods with minimal influence from river regulations and a sufficient quality for high streamflows; see Engeland et al. (2016) for details.

The flood data used in this study is the median annual maximum flood [l/s/km2] at each station. We used the 1, 6, 12, 18, 24, 36, and 48 hour median annual maximum flood. The data at different durations were constructed via an aggregation-based approach, where the durations represent the total volume of water that arrives over a time span of, for example, 24 hours, not flood events that lasted precisely 24 hours. This approach is used in, for example, Breinl et al. (2021) and Barna et al. (2023). For each station, even spacing in the streamflow time series was enforced via regular sampling of a linear interpolation of the

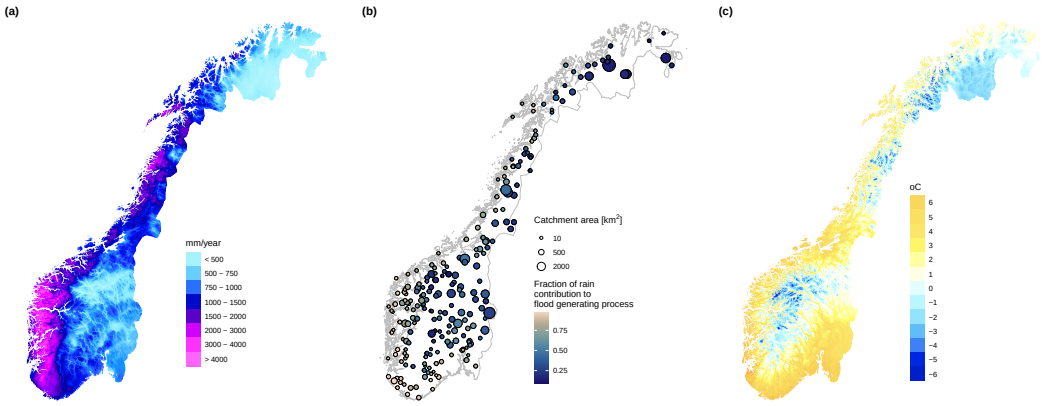

Figure 1. Panel (a) shows average rainfall totals (mm) for the entire year from the period 1991-2020. Panel (b) shows locations of the 232 gauging stations used in this study, where catchment area and average fraction of rain contribution to flood are indicated by size and color, respectively. Panel (c) shows average temperature (°C) for the entire year from the period 1991-2020.

observed data. A moving average with a window length corresponding to the duration of interest was then applied to the evenly spaced streamflow time series. From the smoothed time series, annual maxima were extracted to create separate sets of maxima for each of the durations of interest. We then computed the median of these sets of maxima to get the 1, 6, 12, 18, 24, 36, and 48 hour median annual maximum flood at each of the 232 stations.

Data quality control

Given the focus on sub-daily floods, it is necessary to make sure that the sampling frequency of the data is high enough to represent peak flood magnitudes with sufficient quality. Each of the streamflow records contains a variety of collection methods. These differing collection methods provide data at different frequencies. Generally, the earlier part of the streamflow record has daily time resolution, while the later part of the record contains a higher frequency of measurements after adoption of digitized limnigraph records and/or digital measurements. For our dataset, the shift to a higher frequency of measurements is typically around 1980, and stations have, on average, 27 years of high frequency data. The time resolution of the digital measurements and the digitization of the limnigraph records were selected by NVE to be frequent enough to represent flood peaks at individual stations. Total record lengths in our data set range from a minimum of 20 years of data to 129 years at station 62.5 (Bulken); the distribution of total record lengths is plotted in Fig. 2.

The median annual maximum flood at each duration is computed over the total number of years of data available at each station. This means that for certain stations, especially those with longer record lengths, the median is constructed from annual maxima derived from streamflow time series at a combination of different resolutions. Thus it is of interest to know what percentage of the record is comprised of subdaily data. We calculate the number of years of subdaily data for each station as all years that have at least 200 days of subdaily data. Figure 2 shows the distribution of the subdaily record percentage in our dataset. Around 100 stations have subdaily data percentages over 90 %. The other stations have percentages of subdaily

data that range from 20 % to 90 %. Any station that has less than half of its record made up of subdaily data was manually validated to ensure that the sampling frequency adequately captured flood peaks at those locations. The stations showing a low percentage of subdaily data are characterized by having a long total record length compared to the subdaily record length, i.e. in these cases, the amount of subdaily data is not below average; rather, the overall record length is extensive. There was no correlation between model performance at the 1-hour duration and either total record length or percentage of the record that was subdaily data for each of the model evaluation metrics used in this study (results not shown).

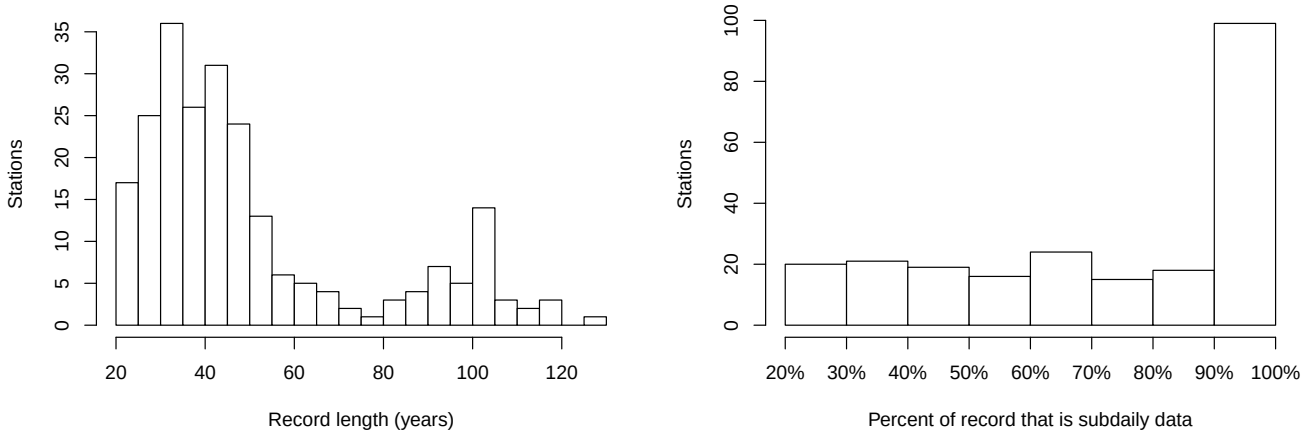

Figure 2. Histograms for record length (years) and percent of the record that is subdaily data. Only years that had at least 200 days of subdaily data count towards the subdaily data total when calculating the record percentage. Stations with less than 50 % of the record comprised of subdaily data were manually validated to make sure the sampling frequency of the data was high enough to represent flood peaks at that location.

In addition to quality control on the sampling frequency, the data have undergone a detailed quality control by the hydrometric section at NVE. Ice jams pose a challenge at numerous stations in Norway and can affect the accuracy of the rating curves used to estimate streamflows from water level measurements. In such cases, specific correction procedures outlined in NVE's internal quality assurance protocols have been implemented to obtain accurate discharge values. Any year with less than 300 days of data was excluded from the analysis.

Catchment descriptors

In addition to streamflow, we derived a set of geographical and hydro-climatic catchment descriptors for each catchment. Table 1 lists the 76 total catchment descriptors. The geographical descriptors include size and shape of the catchments, length of the river networks, land use properties and elevation distributions. The hydro-climatic descriptors include mean annual runoff, and mean annual and monthly precipitation, temperature and sum of rain and snow melt. All these descriptors are available on a 1x1 km grid covering all Norway. The mean annual runoff was based on the runoff map for Norway (Beldring et al., 2002), whereas all other hydro-climatic descriptors were based on interpolated observations given by the SeNorge 2.0 dataset

(Lussana et al., 2019). Rain was defined as precipitation when the temperature is above 0 °C. Snow melt was extracted from the SeNorge snow model (Saloranta, 2014).

For this study, we determined the average contribution of rainfall to floods at each catchment by calculating the ratio of rainfall to the total water depth, where the total water depth includes both rainfall and snowmelt accumulated within a specific time period prior to each flood. These ratios were then averaged across all flood events for the catchment. For details see Engeland et al. (2020).

The catchment areas vary from 0.52 km^2 to 6182 km^2, with a median size of 124 km^2. Roughly half (53 %) of the catchments have more than 1 % of their area covered by lakes; of these catchments, the median effective lake percentage is 2.8 %. Mean annual precipitation ranges from 390 mm to 3196 mm, displaying a notable east-west gradient across the country, with higher precipitation levels along the west coast. The mean annual temperature ranges from -4.0 °C to 7.2 °C, with a median of 0.15 °C. Temperature is influenced by both elevation and latitude; temperature decreases as elevation and latitude increase. The minimum altitude of the catchments spans from sea level to 1104 m.a.s.l., while the catchment relief varies from 54 m to 2019 m. Catchments with greater relief are typically located in the mountain ranges along the west coast, which exhibits more rugged topography than the flatter regions of the country in the east.

STUDY DESIGN

A flowchart of the study design is displayed in Fig. 3. Panel (a) details the process of predictor selection for the GAM developed in this study (*floodGAM*). The predictor selection process is divided into two parts. The first part (Part I, identification of promising predictors) is complementary to, but not necessary for, the second part. We describe our approach for identifying promising predictors for expert assessment in the first subsection of the methods section, and note that other variable identification techniques or expert judgement alone could replace the approach detailed in this section. Part II, selection of predictors for floodGAM, is described in the following subsection. Panel (b) details the validation and visualization process for floodGAM. In the validation step, floodGAM is compared to the two benchmark models, RFFA_2018 (the existing log-linear model for median flood estimation in Norway) and XGBoost. The benchmark models are summarized in the methods section. We assess the performance of the models through a cross-validation study, such that predictive accuracy and reliability are assessed through the consistency between predictions and holdout data. Predictive performance for floodGAM and RFFA_2018 is assessed on five evaluation metrics. XGBoost provides a supplementary benchmark value for the mean absolute error (MAE); due to distributional assumptions, we cannot obtain optimal predictors for XGBoost for the other four evaluation metrics. Reliability is assessed for the 1 and 24 hour durations through the probability integral transform (PIT) which is also only available for RFFA_2018 and floodGAM. In order to keep the results section concise, visualization and reliability metrics are reported only for the 1 and 24 hour durations. These are the durations most relevant to flood guidelines in Norway. Predictive performance results are reported in the first subsection of the results section. Relibability results are reported in the second subsection. Finally, the third subsection of the results section presents the visualization and comparison of the data-driven relationships between predictors and the response identified by floodGAM between the 1 and 24 hour durations.

Table 1. Descriptions of the 76 catchment descriptors used in the study, grouped into geographical and hydro-climatic descriptors. Abbreviations are further used in the text and figures.

Variable	Description	Unit
A	Logarithm of catchment area	km^2
O	Catchment circumference	m
A_P	Catchment area / circumference * 1000	km
D, D_{net}	Drainage density (total river length / area), (total river length excluding lakes / area)	-
C_L	Logarithm of catchment length	km
R_L	Length of main river	km
R_{TL}, R_L	Total river length, and total river length excluding lakes	km
R_G, R_{G1085}	Gradient of main river, and gradient of main river excluding the 10 % lowest and 15 % highest reaches	m/km
H_{10}, H_{50}, H_{90}, H_{MAX}, H_{MIN}	The 10th, 50th, and 90th percentile of the hypsographic curve, maximum elevation, minimum elevation	m.a.s.l.
H_F	Catchment relief (maximum elevation - minimum elevation)	m
C_S	Mean slope	°
A_{Glac}, A_{Agr}, A_{Bog}, A_U, A_L, A_{LE}, A_{For}, A_{Mount}	Percentage of catchment covered by glaciers, agriculture, bogs, urban areas, lakes, effective lake percentage, forests, mountains	%
Q_N	Mean annual runoff 1961-1990	$l/s/km^2$
P_{Jan}, P_{Feb}, P_{Mar}, P_{Apr}, P_{Mai}, P_{Jun}, P_{Jul}, P_{Aug}, P_{Sep}, P_{Oct}, P_{Nov}, P_{Dec}	Mean precipitation from 1961-1990 in January, February, March, April, May, June, July, August, September, October, November, December	mm/month
P_N	Mean annual precipitation 1961-1990	mm/year
$P_{Med1Max}$, $P_{Med2Max}$, $P_{Med3Max}$, $P_{Med4Max}$, $P_{Med5Max}$	Median of annual 1-, 2-, 3-, 4-, and 5-day precipitation	mm/day
T_{Jan}, T_{Feb}, T_{Mar}, T_{Apr}, T_{Mai}, T_{Jun}, T_{Jul}, T_{Aug}, T_{Sep}, T_{Oct}, T_{Nov}, T_{Dec}	Mean temperature from 1961-1990 in January, February, March, April, May, June, July, August, September, October, November, December	°C
T_N	Mean annual temperature 1961-1990	°C
W_{Jan}, W_{Feb}, W_{Mar}, W_{Apr}, W_{Mai}, W_{Jun}, W_{Jul}, W_{Aug}, W_{Sep}, W_{Oct}, W_{Nov}, W_{Dec}	Mean sum of rainfall and snowmelt from 1961-1990 in January, February, March, April, May, June, July, August, September, October, November, December	mm/month
W_N	Mean annual sum of rainfall and snowmelt 1961-1990	mm/year
$W_{Med1Max}$, $W_{Med2Max}$, $W_{Med3Max}$, $W_{Med4Max}$, $W_{Med5Max}$	Median of annual 1-, 2-, 3-, 4-, and 5-day rainfall and snowmelt	mm/day

METHODS

Identification of promising predictors using a machine learning-based algorithm

The algorithm used to identify catchment descriptors that are potentially useful in predicting the median flood is the Iterative Input Selection (IIS) algorithm proposed in Galelli and Castelletti (2013). Given a large set of potentially co-linear predictors, the IIS algorithm selects a smaller, non-redundant set of predictors using a ranking procedure and a stepwise forward selection process. Candidate variables are ranked using an input ranking algorithm, and the top-ranked variables are evaluated by adding them to the selected variable set and measuring prediction accuracy on a chosen model. This process is repeated with residuals as the new response variable until the best variable is already in the set or the model's performance does not improve. The algorithm in full can be found in Galelli and Castelletti (2013).

IIS requires the choice of (i) an input ranking algorithm and (ii) a model to evaluate the predictive performance of the chosen subset of candidate variables. A tree-based ensemble is an effective choice for both (i) and (ii) since the ensemble can be directly exploited as an input-ranking procedure; the structure of tree-based ensembles can be used to infer the relative

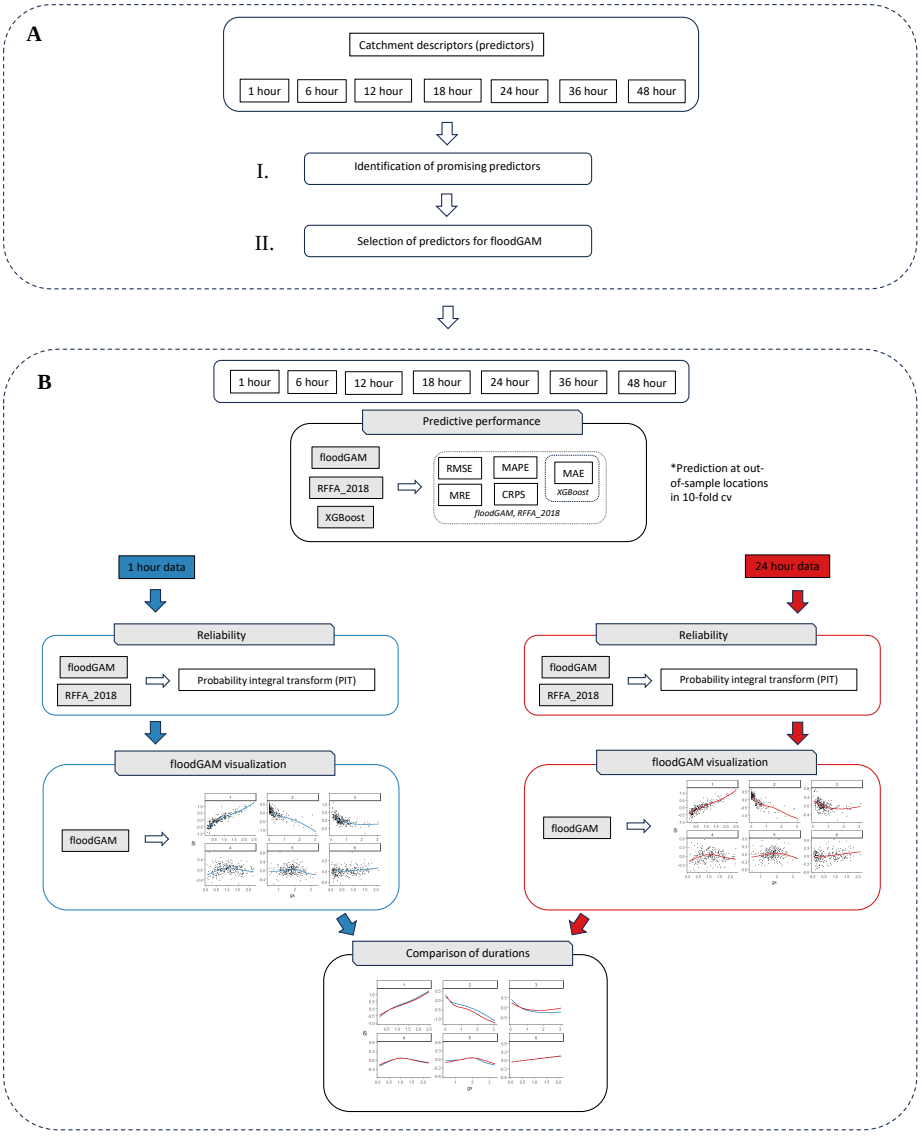

Figure 3. Flowchart of the study design. Panel (a) details the process of predictor selection for the GAM developed in this study. Panel (b) details the model validation and visualization process.

importance of input variables and order them accordingly. Galelli and Castelletti (2013) use a tree-based ensemble for both the input ranking algorithm and predictive model. We do the same, except in our study we must additionally ensure that the relationships generated by the tree-based ensemble are simple enough to be able to be modeled by the GAM. To account for

this, the tree depth—an important parameter controlling the interaction depth between input variables—is restricted to one, since we do not consider variable interactions in the GAM.

We choose XGBoost as the tree-based ensemble within IIS in line with the recent work of Alsahaf et al. (2022). XGBoost is a popular open-source software implementation of extreme gradient tree boosting (Chen et al., 2015; Chen and Guestrin, 2016), which was first proposed in Friedman et al. (2000) and is a computationally efficient implementation of the gradient tree boosting from Friedman (2001). Details of the algorithm set up and hyperparameter tuning for XGBoost can be found in Appendix .

In this study, we run the IIS algorithm within a resampling method to assess consistency of the selected variable sets. This is important in context of flagging promising predictors as there is no uncertainty associated with the XGBoost output or the selected variable sets from IIS. For the resampling step in this study, we choose to systematically resample without replacement, splitting our data into ten non-overlapping folds; however, we note that other resampling methods, such as bootstrap, could also be used as the resampling step. This repeated application of IIS to subsampled data means each application of the algorithm could potentially select a different variable set, where both the chosen variables and the total number of variables are allowed to vary. A visual explanation of the IIS algorithm within the resampling method can be found in Appendix . The procedure is repeated once for each duration such that we can assess the consistence of selected variables across durations as well as across data folds. In total, the IIS algorithm is applied to 10 subsampled data sets $\times$ 7 durations for a total of 70 applications. The consistency of the selected variable sets is assessed across these 70 applications.

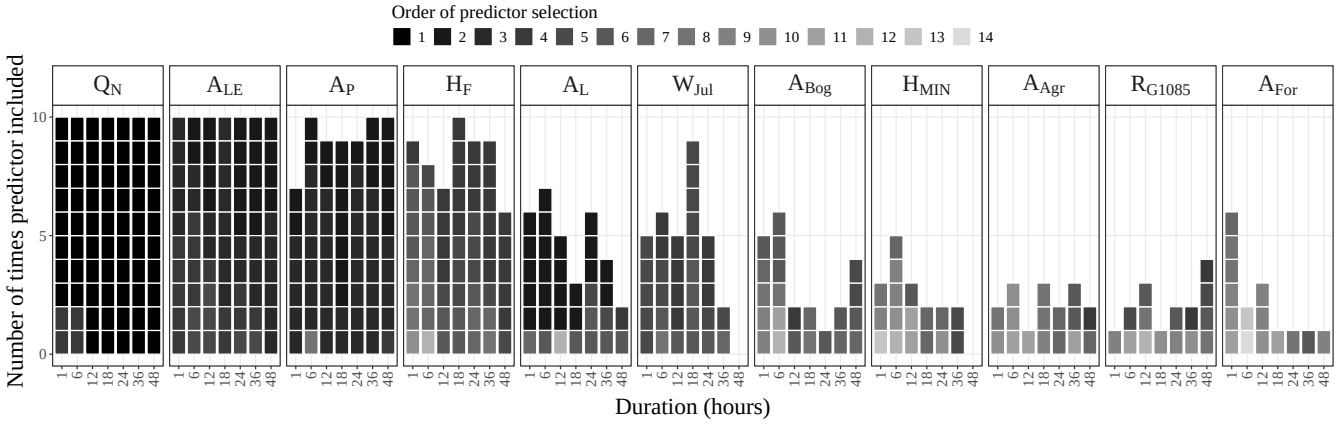

Figure 4. Variables from the pre-selection scheme that appear in at least 15 of the folds. The vertical axis represents the number of times a variable was chosen. The horizontal axis indicates the duration that generated the covariate set. The color indicates the order of variable selection within the IIS algorithm. Variables selected first tend to be those that are most informative.

The IIS algorithm was run with all 76 catchment descriptors as input. Of those 76 variables, 42 were selected by at least one of the folds and 11 were selected by at least fifteen folds. The subset of variables that appeared most consistently is depicted in Fig. 4. The complete set of variables identified by IIS can be found in Appendix . In Fig. 4, the horizontal axis represents the

number of times a variable was chosen within the resampling scheme, out of a maximum possible of 10 (once for each fold). The color of the grid cells represents the order of variable selection within the IIS algorithm. Variables selected first tend to be those that are most informative. For example, the predictor variable Q_N (mean annual runoff from 1961-1990 [l/s/km^2]) consistently emerges as the most important variable across all folds and durations: every fold of subsampled data chooses Q_N as the first predictor. Both consistency of inclusion and order of variable selection can be considered when choosing which predictors to carry forward for more formal variable selection within the model architecture of the GAM.

The catchment descriptors that were most consistently selected include Q_N, A_{LE}, A_P and H_F. The catchment descriptor A_L, which is highly correlated to A_{LE}, is also included for a minority of folds, but its inconsistent inclusion sets it apart from A_{LE}. A similar situation arises for the predictors H_F and H_{MIN}. The other less consistently selected predictors are W_{Jul}, which describes the mean sum of rain and snow melt in July from 1961-1990 [mm/month]; A_{Bog}, the percentage of the catchment covered by bogs; H_{MIN}, the minimum elevation of the catchment [meters above sea level]; A_{Agr}, the percentage of the catchment used for agriculture; R_{G1085}, which describes the gradient of main river excluding the 10 % lowest- and the 15 % highest reaches [m/km]; and A_{For}, the percentage of the catchment covered by forests. The different durations generally selected the same predictors, particularly on those that were most consistently selected (Q_N, A_{LE}, and A_P). Duration specific differences beyond this should not be over interpreted given the variability in the full selected set shown in Appendix .

Predictor selection for floodGAM

The predictors chosen for floodGAM are identified in Table 2 and include four geographical catchment descriptors: (1) A_{LE} - effective lake percentage, (2) A_P - the area of a catchment divided by its circumference, (3) R_{G1085} - the gradient of the main river excluding the 10 % lowest and 15 % highest reaches, (4) H_F - the difference in catchment elevation from the highest to lowest point, as well as three hydro-climatic catchment descriptors: (5) Q_N - the mean annual runoff, (6) W_{Apr} - the mean sum of rainfall and snowmelt in April, and (7) P_{Sep} - the mean precipitation in September. These predictors were chosen using the results from the IIS algorithm (Fig. 4) in combination with expert judgement. For example, the climate descriptor P_{Sep} was not chosen by IIS but was added because the autumn- and winter flood season–with mainly rainfall-driven floods–is important in Norway and P_{Sep} is a good representation of the autumn precipitation. Similarly, W_{Apr} is chosen over W_{Jul} as expert judgement found it more informative to include information on the spring flood season. Land use and land type predictors (A_{Agr}, A_{Bog}, A_{For}) were less consistently selected by the IIS algorithm and were excluded from the final predictor set based on expert judgement and experience with the data set. The predictor R_{G1085} is heavily skewed; we found it useful to log transform that predictor for a more numerically stable estimation within floodGAM. All seven predictors were verified as significant by shrinkage estimation within the implementation of floodGAM.

One of the benchmark models, RFFA_2018, is the log-linear model currently used by NVE to predict the median flood (Engeland et al., 2020). The RFFA_2018 model was developed for 24 hour flood data. Table 2 displays the catchment descriptors, and their transformations, used in RFFA_2018 and floodGAM. Predictors and predictor transforms in RFFA_2018 were chosen according to internal protocols at the Norwegian Water and Energy Directorate. The other benchmark model, XGBoost, has access to all 76 catchment descriptors. Previous research (Alsahaf et al., 2022) observed improved predictive performance

with XGBoost models when employing IIS-based pre-selection; however, our analysis found that pre-selection applied to the XGBoost models in this study did not alter the statistical significance of the results. For simplicity, all reported intervals and evaluation metrics pertain to the XGBoost model applied to the full catchment descriptor set.

Table 2. Descriptions of predictors used in the models floodGAM and RFFA_2018, structured into geographical (top) and hydro-climatic (bottom) catchment descriptors. Abbreviations are further used in figures. Inclusion of predictor variables is indicated for each model, and variable transformations are listed in their respective rows.

Variable	Description	floodGAM	RFFA_2018
R_L	Length of main river [km]		sqrt(x)
A_{LE}	Effective lake percentages [%]	x	x
A_P	Catchment area / circumference * 1000 [km]	x	
R_{G1085}	Gradient of main river excluding the 10 % lowest- and the 15 % highest reaches [m/km]	log(x)	
H_F	Maximum elevation - minimum elevation [m]	x	
Q_N	Mean annual runoff 1961-1990 [l/s/km^2]	x	x$^{1/3}$
T_{Feb}	Mean temperature February 1961-1990 [°C]		x^2
T_{Mar}	Mean temperature March 1961-1990 [°C]		x^3
W_{Mai}	Mean sum of rain and snow melt May 1961-1990 [mm/month]		sqrt(x)
W_{Apr}	Mean sum of rain and snow melt April 1961-1990 [mm/month]	x	
P_{Sep}	Mean precipitation September 1961-1990 [mm/month]	x	

Generalized Additive Models

GAMs, introduced by Hastie and Tibshirani (1987), are a class of regression models that extend the linear regression model to handle non-linear relationships between the predictor variables and the response variable. GAMs model the relationship between the response variable and each predictor separately by assuming a smooth, continuous, non-parameteric function of each predictor. These functions are then combined additively to obtain the overall prediction. This allows for a wide range of predictor-response relationships to be captured without specifying a prior functional form. Furthermore, these predictor-response relationships are easily visualized by plotting the partial response curve for each predictor.

Let $\mathbf{y}$ be a vector of our response variable (the median flood at location i) with index $i \in [1, \ldots, n]$ referring to the ith element. Then the GAM relates the mean response for observation i to the sum of smooth functions of p explanatory variables $x_{i1}, \ldots, x_{ip}$ as follows:

$$g\left(\mathbf{E}[y_i]\right) = \alpha + \sum_{j=1}^{p} s_j\left(x_{ij}\right) \tag{1}$$

where $s_j()$ is the smooth function of predictor x_{ij}, α is the intercept and $g()$ is a monotonically differentiable link function. The smooth function $s_j()$ is defined by a linear combination of basis functions, allowing the relationship between x_{ij} and

response y_i to be non-parametrically modeled. Because this non-parametric construction is so flexible, selecting the appropriate level of 'smoothness' for each predictor is an important component of GAM construction. In practice, this is often done by limiting the effective degrees of freedom. We used a thin plate spline basis with effective degrees of freedom limited between 6 and 3 for our chosen predictors.

While the form of the predictor-response relationship can be modeled non-parametrically, the probability distribution of the response variable in the GAM must still be specified. We chose to model the data as normally distributed with a log link, in line with standard practices in hydrology that model flood volumes and flood peak discharges as log normal (Stedinger, 1980).

We used the 'mgcv' package in the R statistical software (Wood, 2017) to implement the GAMs. The mgcv package contains a convenient variable selection method based on null-space penalization, which allows smooth functions associated with a particular predictor to be penalized to the zero function and thereby selected out of the model if the predictor is nonimportant (Marra and Wood, 2011). This capability is accessed by setting the *select* argument of the gam() function to 'True'. We set select = T and use restricted maximum likelihood ('REML') as the estimation method for each of the GAMs in this study.

Benchmark models

RFFA_2018

The existing model for index flood estimation in Norway (RFFA_2018) is the log-linear model presented in Engeland et al. (2020). Let $\mathbf{y}$ be a vector of our response variable with index $i \in [1, \ldots, n]$ referring to the ith element. Let $\mathbf{X}$ be our predictor matrix with $n \times p$ elements, where p is the number of predictor variables. Furthermore, since we wish to evaluate the predictor-response relationship on the log scale, let $z_i = \log(y_i)$. Then the regression equation is given as

$$z_i = \alpha + \sum_{j=1}^{p} \beta_j x_{ij} + \epsilon_i, \tag{2}$$

where α and β_j for $j = 1, \ldots, p$ are the parameters to be estimated and where the ϵ_i are mutually independent error terms assumed normally distributed $N(0, \sigma^2)$. The mgcv package provides routines to fit log-linear models as well as GAMs and was used to estimate RFFA_2018 in this study.

XGBoost

XGBoost is a popular open-source software implementation of extreme gradient tree boosting (Chen et al., 2015; Chen and Guestrin, 2016), which was first proposed in Friedman et al. (2000) and is a computationally efficient implementation of the gradient tree boosting from Friedman (2001).

Gradient tree boosting is a machine learning technique that involves training an ensemble of decision trees sequentially, with each subsequent tree aimed at reducing the residual errors of the previous tree. At each step, a gradient descent algorithm is used to optimize a predefined loss function by adjusting the weights of the features in each tree.

Let $\mathbf{y}$ be a vector of our response variable with index $i \in [1, \ldots, n]$ referring to the ith element. Let $\mathbf{X}$ be our predictor matrix with $n \times p$ elements, where p is the number of predictor variables. Furthermore, since we wish to evaluate the predictor-response

relationship on the log scale, let $z_i = \log(y_i)$. Then the regression equation is given as

$$\hat{z}_i = \sum_{k=1}^{K} f_k(\mathbf{x}_i),\ f_k \in \mathcal{F}, \tag{3}$$

where f_k, $k \in [1, \ldots, K]$ is the ensemble of regression trees and K is the number of trees used. Here $\mathbf{x}_i$ is the ith row of the predictor matrix and $\mathcal{F}$ is the set of all possible classification and regression trees (e.g. CARTs; see XGBoost documentation). Then the objective function to be minimized is given by

$$L^k = \sum_{i=1}^{n} L\left(z_i, \hat{z}_i^{k-1} + \eta f_k(\mathbf{x}_i)\right) + \Omega(f_k) \tag{4}$$

where L^k is the kth iteration loss, $\hat{z}_i^{k-1}$ is the prediction at the previous iteration, η is a shrinkage parameter controlling the learning rate, f_k is the tree that provides the best improvement to the model as measured by the predefined loss function, and $\Omega(f_k)$ is a penalization parameter that controls the complexity of trees to avoid overfitting. Here we used the squared error loss as the objective function:

$$L = \sum_{i=1}^{n} (z_i - \hat{z}_i)^2. \tag{5}$$

The following hyperparameters were tuned on the indicated ranges: tree depth (1-10); the percentage of observations subsampled at each boosting step (0.1-1); the minimum number of instances needed in each node (1-5); and the shrinkage parameter η (0.01-0.1). The number of boosting iterations was evaluated up to a maximum number of 999 iterations. Hyperparameter tuning was conducted within a 10-fold cross-validation scheme using all possible parameter combinations and an early stopping criterion for the number of boosting iterations, where the algorithm stopped after 25 rounds without improvement in the error rate. The ranges of the hyperparameters were chosen based on experience with the data set and recommended XGBoost practices.

Evaluation methods

This section presents (i) the error metrics used to evaluate the predictive performance of the models, (ii) a computationally efficient permutation test that allows us to assess the statistical significance of differences in error metrics between the models (Thorarinsdottir et al., 2020) and (iii) the probability integral transform (PIT). The PIT is used to assess reliability of the models as measured by the consistency between model predictions and validation data. We assess the performance of the models through a cross-validation study, such that predictive accuracy and reliability are assessed through the consistency between predictions and holdout data.

Error metrics

We evaluate model performance using the root mean squared error (RMSE), the mean absolute error (MAE), the mean relative error (MRE), the mean absolute percent error (MAPE) and the continuous ranked probability score (CRPS) (Gneiting and

Raftery, 2007; Hersbach, 2000). All of these metrics measure slightly different aspects of the predictive distribution. The RMSE, MAE, and CRPS are expressed in the units of the response variable (l/s/km2) and give more weight to catchments with higher discharge values. In our case, these metrics tend to prioritize minimizing errors in catchments located on the west coast of Norway, where the median flood values, given in [l/s/km^2], are the highest. The proportional error metrics–the MAPE and the MRE–avoid this issue of scale but are sensitive to highly over- or under-estimated values. Four of the metrics here (RMSE, MAE, MRE, MAPE) assess the distance between an observed value and a single predicted value; that is, they are error metrics for point forecasts. The CRPS measures the difference between the predicted and observed cumulative distributions (Hersbach, 2000) and thus provides a measure of how variable the predictions are in addition to assessing accuracy.

Constructing a statistically meaningful model ranking from these error metrics requires that the predicted value from the model minimizes the given error metric. For example, the root mean squared error (RMSE) is minimized when the predicted value is chosen as the mean of the predictive distribution. If an alternative distributional feature, such as the median, is used with the RMSE in the situation where the median and mean of the predictive distribution are not equivalent (e.g., when the data are assumed log normal), any model rankings constructed from the resulting quantity will be unreliable.

We list the optimal predictor (minimizing quantity) for each of the error metrics for point forecasts in Table 3. The MRE and MAPE are minimized by the functionals given in Gneiting (2011); see Appendix for calculation of the optimal predictors here. Both floodGAM and RFFA_2018 are assessed on all five metrics. XGBoost is assessed only on the MAE; optimal predictors for the other four error metrics are not accessible for XGBoost when the data are assumed log normal. Table 3 defines the metrics and reports the associated units. All metrics are negatively oriented, i.e. a smaller value indicates better predictive performance.

Table 3. Definitions of error metrics used in this study: root mean squared error, mean absolute error, mean relative error, mean absolute percent error, and continuous ranked probability score. Here $\hat{y}_i$ is the predicted value at station i, $i \in [1, \ldots, n]$, y_i is the observed value at station i, and F_i is the cumulative distribution function of the predictive distribution with finite first moment. For the CRPS, $H(y - y_i)$ denotes the Heaviside function and takes the value 0 when $y < y_i$ and the value 1 otherwise.

Error metric		Optimal predictor	Units
RMSE	$\sqrt{\left(\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2\right)}$	$\hat{y}_i = \text{mean}(F_i)$	l/s/km2
MAE	$\frac{1}{n}\sum_{i=1}^{n}\lvert y_i - \hat{y}_i\rvert$	$\hat{y}_i = \text{median}(F_i)$	l/s/km2
MAPE	$\frac{1}{n}\sum_{i=1}^{n}\left\lvert\frac{y_i - \hat{y}_i}{y_i}\right\rvert \cdot 100$	$\hat{y}_i = \text{med}^{(-1)}(F_i)$	%
MRE	$\frac{1}{n}\sum_{i=1}^{n}\left\lvert\frac{y_i - \hat{y}_i}{\hat{y}_i}\right\rvert \cdot 100$	$\hat{y}_i = \text{med}^{(1)}(F_i)$	%
CRPS	$\int_{-\infty}^{\infty}\left[F_i(y) - H(y - y_i)\right]^2 dy$	-	l/s/km2

Permutation test

The permutation test, see e.g. Thorarinsdottir et al. (2020), determines the difference in scores between two models, A and B, by computing

$$c = \frac{1}{n}\sum_{i=1}^{n}(\phi(A_i, y_i) - \phi(B_i, y_i)) \tag{6}$$

Here, n represents the total number of stations and $\phi(\cdot)$ is an error metric from Table 3. The metric takes the observed value y_i and the model output A_i or B_i at station i as inputs. If c is negative, it indicates that model A performs better than model B in terms of the error metric, and vice versa. The permutation test creates resampled copies of c with randomly swapped model index A and B for each i. Under the null hypothesis that both models perform equally well, the set of permutations cannot be differentiated from c. The statistical test formalizes this concept by determining which quantile c occupies in the set of permutations; if the p-value is less than 0.05, then it suggests that the performance of model A is significantly better than model B.

Probability integral transform

Reliability describes the consistency between model predictions and validation data. A reliable model is expected to produce an estimated distribution that closely aligns with the unknown true distribution of the data. This is typically assessed through the probability integral transform (PIT); if the observations follow the estimated distribution, the PIT values will be approximately uniformly distributed (Dawid, 1984):

$$F_i(y_i) \dot{\sim} U([0,1]).$$

The uniformity of the PIT values is represented graphically through histograms. As the calculation of the PIT values requires a cumulative distribution function, F_i, the reliability assessment is not accessible for the XGBoost models in this study.

RESULTS

Predictive performance

We report the performance evaluation metrics for floodGAM, RFFA_2018, and XGBoost in Table 4. The best result is shown in bold font. If floodGAM was statistically significantly better than RFFA_2018 at the $\alpha = 0.05$ level on a particular metric and duration, the significance is indicated with an asterisk. The predictive performance for floodGAM predicting across durations for the 1 and 24 hour durations–that is, using floodGAM fit on the 24 hour data to predict at the 1 hour duration and vice versa–is shown and the duration used to fit the model is indicated in the model name ("floodGAM, 24 hours" and "floodGAM, 1 hour").

On the 1 hour duration, floodGAM was statistically significantly better than RFFA_2018 on all error metrics. It was also statistically significantly better than RFFA_2018 for certain error metrics on the 6 and 12 hour durations. There were no statisti-

cally significant differences between floodGAM and RFFA_2018 at durations longer than 12 hours, or between floodGAM and XGBoost on the MAE at any duration. The floodGAM predicting across durations was not competitive and was statistically significantly worse than the duration-specific floodGAM.

To illustrate how floodGAM improves on RFFA_2018 at the 1 hour duration, we plot a model-by-model comparison of the error at each station (Fig. 5). Here we show three different error metrics (absolute percent error, relative error and the CRPS). Figures for other metrics, models and durations can be found in Appendix . Points falling above the diagonal line indicate stations where RFFA_2018 performed worse than floodGAM. Points falling below the diagonal line indicate stations where floodGAM performed worse than RFFA_2018. Point size shows catchment area, point color indicates the fraction of rain contribution to flood.

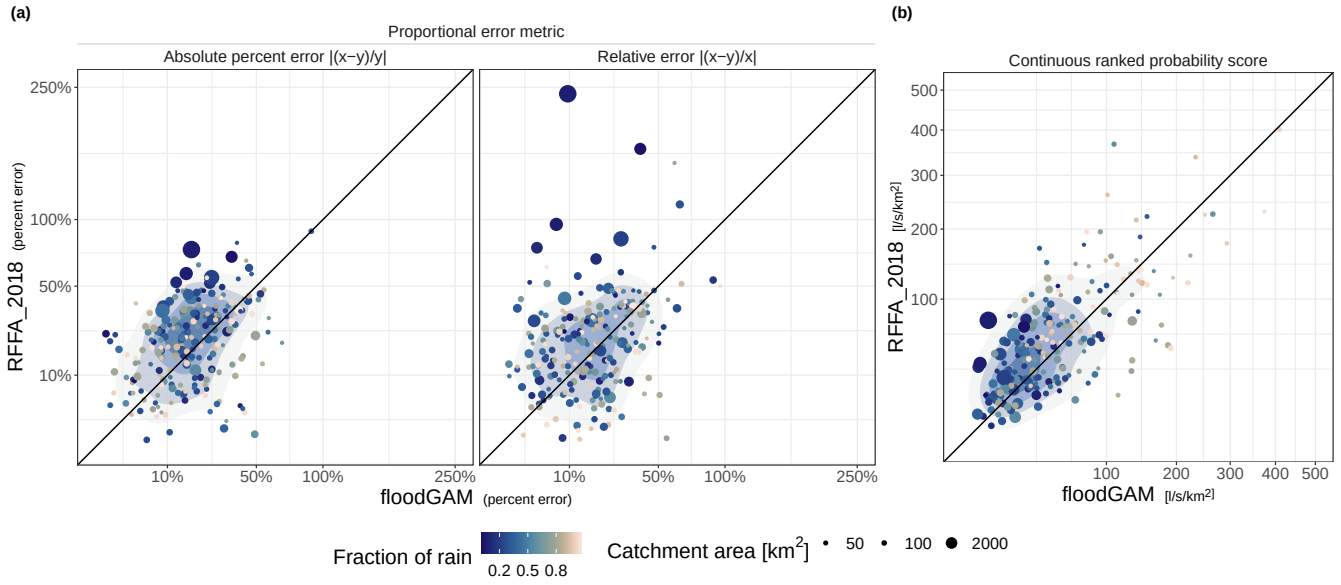

Figure 5. Model to model comparison on absolute percent error, relative error, and the continuous ranked probability score for RFFA_2018 and floodGAM on the 1 hour duration. In the panel headers, x represents the predicted value and y the observed value. Points falling above the diagonal line indicate stations where RFFA_2018 performed worse than floodGAM. Points falling below the diagonal line indicate stations where floodGAM performed worse than RFFA_2018. The 2D kernel density estimation of point density is underlaid to aid visual interpretation. Point size shows catchment area, point color indicates the fraction of rain contribution to flood.

Figure 5 shows that RFFA_2018 systematically underestimates the 1 hour median flood in large, snowmelt driven catchments (Panel (a) of Fig. 5). The fact that these large, snowmelt driven catchments have relative errors that are greater than the absolute percent error means that the observed values are higher than the predicted values: RFFA_2018 is underestimating at these stations. This underestimation is not obvious when looking at the absolute percent error as the absolute percent error supports severe underestimation (Gneiting, 2011). The opposite effect–supporting severe overestimation–is true for the relative error. In addition to improved performance in extreme cases, Fig. 5 shows that floodGAM has better performance in the bulk of the

data; that is, there is a higher density of points above the diagonal. This is visualized through the kernel density estimation underlaid on each panel (shaded areas in Fig. 5). Panel (b) shows the catchments with high CRPS values are typically rain-driven catchments with small area and large discharge values.

Table 4. Model evaluation metrics–root mean squared error, (mean) continuous ranked probability score, mean absolute error, mean relative error, mean absolute percent error–showing predictive performance for floodGAM and the benchmark models. The best result is shown in bold font. If floodGAM was statistically significantly better at the $\alpha = 0.05$ level than RFFA_2018 on a particular metric and duration, the significance is indicated with an asterisk.

Duration	Name	Evaluation metric				
		RMSE [l/s/km2]	CRPS [l/s/km2]	MAE [l/s/km2]	MRE [%]	MAPE [%]
1 hour	floodGAM	**118.8***	**59.8***	**84.5***	**19.8***	**20.3***
	floodGAM, 24 hours	185.6	82.0	111.3	25.1	22.4
	RFFA_2018	136.1	69.0	95.4	25.2	24.0
	XGBoost	-	-	91.6	-	-
6 hour	floodGAM	**112.1**	**55.9***	**79.0**	**18.8***	**19.6***
	RFFA_2018	121.7	61.5	85.9	22.4	22.1
	XGBoost	-	-	85.1	-	-
12 hour	floodGAM	107.0	**51.6**	**71.9**	**18.0***	**18.7**
	RFFA_2018	**106.0**	53.4	76.0	19.8	19.9
	XGBoost	-	-	74.9	-	-
18 hour	floodGAM	94.9	**46.8**	**65.3**	**17.5**	**18.0**
	RFFA_2018	**93.3**	47.3	67.0	18.2	18.4
	XGBoost	-	-	67.4	-	-
24 hours	floodGAM	85.8	**43.3**	**60.6**	17.1	17.8
	floodGAM, 1 hour	163.0	75.8	106.7	23.5	25.6
	RFFA_2018	**85.4**	43.4	61.3	17.1	**17.4**
	XGBoost	-	-	62.8	-	-
36 hour	floodGAM	72.2	37.3	52.8	16.1	16.9
	RFFA_2018	**71.2**	**36.8**	**51.8**	**15.8**	**15.9**
	XGBoost	-	-	53.5	-	-
48 hour	floodGAM	64.8	33.8	47.6	15.7	16.2
	RFFA_2018	**63.5**	**33.1**	**47.0**	**15.3**	**15.3**
	XGBoost	-	-	47.2	-	-

Model reliability

We assess the reliability of the predictions for floodGAM and the existing model, RFFA_2018. Figure 6 shows histograms for floodGAM and RFFA_2018 at the two durations most relevant to flood guidelines in Norway (1 and 24 hour durations). Histograms for both models are roughly uniform but show some evidence of bias: RFFA_2018 has an excess of values at high quantiles, while floodGAM has an excess of values at low quantiles. The bias in RFFA_2018 shows a tendency to underestimate predicted values; this is consistent with results shown by the model evaluation metrics in the previous subsection, where the largest relative errors were caused by underestimations at large, snowmelt driven catchments. From the PIT histogram we see that the bias in floodGAM, on the other hand, tends toward overestimation, although the evaluation metric assessment in Fig. 5 shows none of the overestimations were as severe as the underestimations provided by RFFA_2018.

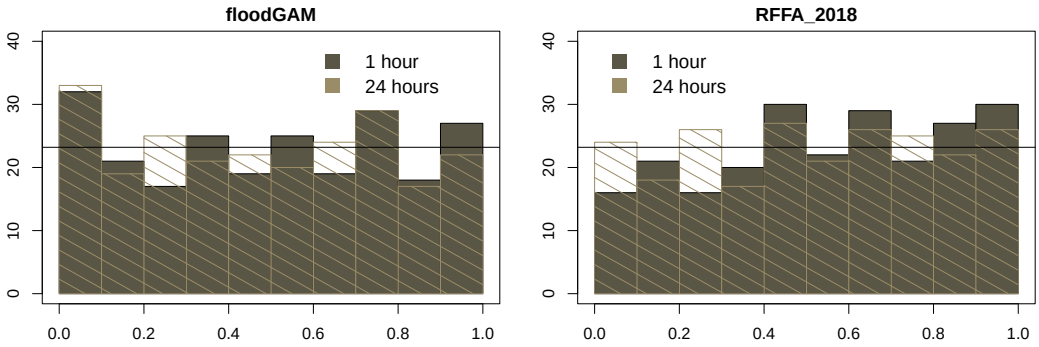

Figure 6. Visualizations of probability integral transform (PIT) values for floodGAM and RFFA_2018 at the 1 and 24 hour durations. If the unknown true distribution of the data is close to the estimated distribution from the models, the PIT values will be uniformly distributed. This is assessed visually via histograms. Both distributions are roughly uniform but show some evidence of bias: RFFA_2018 shows an excess of values at the high quantiles, meaning values tend to be underestimated by this model. The floodGAM model has an excess of values at the low quantiles, indicating a tendency to overestimate.

Table 5 shows the empirical coverage of the associated central 50 %, 80 % and 90 % prediction intervals. The empirical coverage is given by the area under the relevant number of central bins in the histograms in Fig. 6; for example, the 50 % empirical coverage is the total area under the central five bins in the PIT histogram. We replicate that information in numerical form in Table 5 for easy model to model comparison. The average width of the empirical prediction intervals, in l/s/km2, is also shown in Table 5. Wider prediction intervals indicate predictions that are less precise and therefore less informative.

Both models on both durations show empirical coverage probabilities that match the nominal coverage probabilities, reflecting the uniformity we see in the histograms in Fig. 6. However, we see differences in the average width of the probability intervals for the 1 hour duration. For this duration, the RFFA_2018 model reports intervals that are about 30 % wider on average than those for the floodGAM model. This shows the 1 hour predictions for RFFA_2018 are much less precise than those

from floodGAM. The average width of the intervals between the two models is much more similar for the 24 hour duration, although floodGAM still has narrower prediction intervals on average.

Table 5. Empirical coverage and average widths (in l/s/km2) of central prediction intervals for both floodGAM and RFFA_2018. The nominal coverage is 50 %, 80 % and 90 %. Within each duration, models are ordered according to their average CRPS scores (reported in Table 4).

Duration	Model	50 %		80 %		90 %	
		Coverage	Width	Coverage	Width	Coverage	Width
1 hour	floodGAM	48.7 %	143	74.6 %	275	85.8 %	356
	RFFA_2018	52.2 %	189	80.1 %	367	92.7 %	479
24 hours	floodGAM	50.9 %	103	76.2 %	198	87.5 %	256
	RFFA_2018	50.9 %	109	78.5 %	208	90.1 %	270

Explaining the model

The partial response curves for each predictor in floodGAM are plotted in Fig. 7. We display the 1 hour and 24 hour durations. The partial response curves are the smooth components of floodGAM. They show how the median flood varies as a function of a particular predictor when all other predictors are held constant at their mean value. Note that the partial response curves in Fig. 7 are displayed on the link (log) scale, not in the units of the response. Therefore, predictor-response interpretation focuses on whether a predictor has an increasing or decreasing effect on the median flood, and the magnitude of this effect is assessed relative to other predictors in the model. The y axis range of the partial response curves in Fig. 7 indicates the relative importance of the predictors; predictors that are more important have a larger range. Additional information about the relative importance of predictors can be obtained from formal measures such as the likelihood ratios between the full model and the model withholding particular predictors and is displayed in Table 6. Finally, the shading around the partial response curves is the estimation uncertainty associated with each smooth component. Areas with little data–for example, catchments with area to circumference ratios above 5 km–have large estimation uncertainty and possible forms of the smooth component can vary within this uncertainty interval.

Figure 7 also displays the partial residuals associated with each smooth component. Partial residuals for smooth components are the residuals that would be obtained by excluding the specific term from the model while keeping all other estimates fixed. The partial residuals used here are the working residuals from the 24 hour duration added to the corresponding estimate of the smooth term. Coloring the partial residuals by fraction of rain and sizing by catchment area can give a better idea of what types of catchments contribute to the shape of the smooth component. This can aid in identification of predictor-response relationships that are mechanistically realistic.

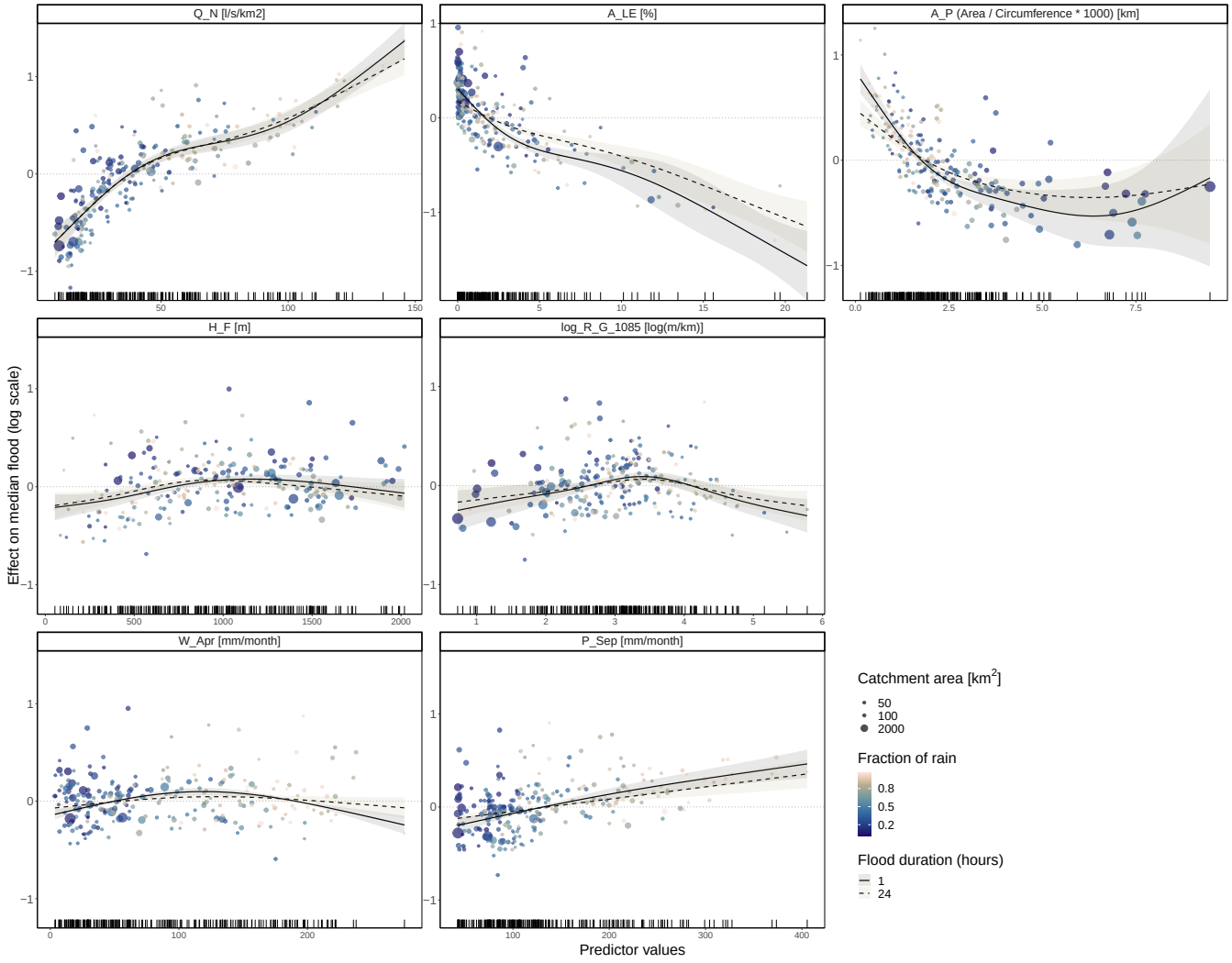

Figure 7. Partial response curves and partial residuals for floodGAM modeling the median annual maxima flood. The smooth components and partial residuals are shown on the link scale, and units for predictors are shown in panel titles. Partial residuals are colored by average fraction of rain contribution and sized by area. The partial residuals used here are the working residuals from the 24 hour duration added to the corresponding estimate of the smooth term. 95 % intervals showing estimation uncertainty for the smooth components are shaded. Location of data points for each predictor are shown as tick marks on the x axis. Y axis ranges span the same magnitude for each panel.

Figure 7 shows that the smooth component for Q_N shows an increasing relationship whereas A_{LE}, A_P and P_{Sep} show a decreasing relationship with the median flood. Three smooth components H_F, R_{G1085} and W_{Apr} have a concave relationship with the median flood.

We see significant differences between durations in the smooth components for A_{LE} and A_P; that is, these smooth components display segments where there is no overlap in the estimation uncertainty intervals between the 1 hour duration and

the 24 hour duration. These two predictors that show duration-specific differences are important predictors. Table 6 reports the predictor ranking by likelihood ratio: A_{LE} ranks as the most important predictor and A_P the third most important predictor. This can also be assessed visually through the y-axis range of the smooth components in Fig. 7 (top row). We do not see significant differences between durations in the smooth component for the second most important predictor, Q_N. Additionally, the remaining four predictors do not show significant duration-specific differences. The least important predictor on both the 1 hour duration and 24 hour duration is W_{Apr}. The other climate variable–P_{Sep}–also has a low importance ranking and is ranked as second to last on the 24 hour duration and third to last on the 1 hour duration. Given the similarity of values between the lowest ranked predictors we do not see the reordering of P_{Sep} between durations as significant.

Table 6. Predictor ranking by likelihood ratio; larger values indicate greater predictor importance. Column "Absolute gain in likelihood value" is given by $-\log(L_0/L_{full})$, i.e. the positive value of the difference in log likelihood between L_0, the model without a particular predictor, and L_{full}, the model with all seven predictors included. The ranking of predictors is consistent between the 1 and 24 hour durations, with the exception of H_F and P_{Sep}.

Predictor	Absolute gain in likelihood value	
	1 hour	24 hours
A_{LE}	129.6	87.7
Q_N	83.7	85.9
A_P	44.2	24.2
R_{G1085}	22.6	15.3
H_F	12.2	12.2
P_{Sep}	12.6	7.7
W_{Apr}	8.6	3.1

When a predictor is correlated with covariates that describe spatial variation of the median flood in Norway, the partial residuals for that predictor show regional groupings (see, for example, Q_N, where the upper end of the smooth effect is dominated by rainfall-driven catchments on the coast of Norway). However, we found no predictive performance benefit when splitting the data into hydrologically homogeneous regions, indicating that the GAM is flexible enough to adjust to differences in response effect between regions within Norway. Evaluation metrics, tests of statistical significance, and regional delineation for the assessment of the GAM on sub-regions are reported in the Appendix.

DISCUSSION

We have, in accordance with our main objective, developed a GAM (floodGAM) such that we could identify and describe the functional relationships between the median flood and catchment descriptors at different durations. Adequacy of floodGAM as an explainable model was established through predictive performance at ungauged locations, where predictive performance

was measured by both predictive accuracy and reliability. The predictive accuracy and reliability of floodGAM matched or exceeded that of the benchmark models at the durations studied.

Hydrologic interpretation of predictor-response relationships in floodGAM

The shape of the smooth components should be interpreted with care: as with all statistical models, there is potential that the relationship reflects unidentified latent or confounding variables rather than a mechanistic relationship with the response. However, we can say the top three most important predictors (Q_N, A_{LE}, and A_P) have smooth components that are consistent with our expectations for the relationship with the median flood. The smooth component for Q_N shows an increasing relationship; as the mean annual runoff for a catchment gets larger so does the median flood. The smooth component for A_{LE} shows a decreasing relationship with the median flood, reflecting the dampening effect of effective lake percentage on flood peak.

The smooth component for A_P shows a decreasing relationship with the median flood. A_P reflects both catchment size and shape. For catchment with similar shapes, A_P increases with catchment area. For catchments with similar areas, A_P is the largest for perfectly circular shapes and the smallest for elongated or irregularly shaped catchments. The circumference used to calculate A_P depends both on the approach used to calculate the catchment boundaries and the resolution of the underlying digital elevation model. In our dataset we assume that the catchment boundaries are consistently defined as they all have a unique source (The Norwegian Water and Energy Directorate, 2010).

For our dataset, the catchment area explains most of the variation in A_P. The decreasing relationship between A_P and median flood can therefore be explained by the well-known spatial scaling of floods (Alexander, 1972; Blöschl and Sivapalan, 1995; Robinson and Sivapalan, 1997a, b; Tsonis et al., 2007; Tarasova et al., 2018; Stein et al., 2021; Najibi and Devineni, 2023). This scaling reflects the changing influence of runoff-generating processes based on catchment size (Blöschl and Sivapalan, 1995; Tarasova et al., 2018), as summarized in Lun et al. (2021): Firstly, a small catchment is more likely to be fully covered by a storm than a large catchment. Consequently, the variance of extreme catchment-average precipitation and thereby the median flood decreases with catchment size (Viglione et al., 2010). Secondly, there is a transition from short-duration convective events to long-duration stratisform precipitation events as the most relevant flood generating process as catchment size increase (Gaál et al., 2015; Merz and Blöschl, 2009). In our data we see also that the snow melt contribution to floods increases with catchment size. Thirdly, the response times increase with area (Gaál et al., 2012) causing smaller flood peaks.

Relationships between catchment shape and flood size is less clear (Stein et al., 2021) and depends on how the time space organization of storm events interacts with the spatial organization of the catchments (Zoccatelli et al., 2011). Based on runoff generation processes, Blöschl (2013) and Viglione and Blöschl (2009) argue that round catchments can be expected to react more quickly than elongated catchments since the flood waves from different parts of the catchment will concentrate quickly. On the other hand, a storm cell that follows a elongated catchments from the top towards the outlet might result in a high flood peak since the flood wave from upstream and downstream parts will overlap (Murthy, 2002). In an empirical study by David and Davidova (2014) the connections between catchment shape and flood magnitude are not significant.

In this study A_P was consistently preferred as a predictor instead of other descriptors reflecting catchment size (A, C_L, R_L, R_{TL}, $R_{TL,net}$), indicating that the catchment shape influences the flood sizes. However, the marginal effect of catchment shape cannot be detected from our model.

Concave relationships between the smooth components H_F, R_{G1085} and W_{Apr} and the median flood are challenging to explain and might be a result of inter-correlated predictors and hidden variables. The smooth component for P_{Sep} shows an increasing linear relationship with the median flood. This is a reasonable relationship for the rainfall-driven catchments that experience high flows during autumn and winter; the partial residuals in Fig. 7 show the increasing nature of the smooth component is driven by catchments with a higher fraction of rain contribution to flood generating process. However, it is less clear that this increasing linear relationship should hold for the snowmelt-driven catchments that experience high flows during spring and summer.

This study was limited to constructing a model for annual maxima since flood guidelines in Norway pertain to annual maximum values; however, as a preliminary investigation into how seasonal flood regimes may influence the shape of the partial response curves shown in Fig. 7, we investigated changes in the partial response curves of floodGAM when seasonal maxima were used instead of annual maxima. Results are reported in the Appendix. We observed season-specific changes in the shape of the partial response curves for climatic variables. These changes were not observed in the partial response curves for the geographical catchment descriptors or the mean annual runoff. This suggests relationships between climatic predictors and annual maxima should be interpreted with caution as these relationships may represent a compromise between different generating processes. This parallels the observations in, for example, Ouarda et al. (2006), McCuen and Beighley (2003), and Fischer and Schumann (2021). Focusing on the role of climatic variables in regression style models that explicitly account for flood generating processes is an interesting area of future research for descriptive statistical studies, particularly when investigating models that incorporate non-stationarities in climate: extrapolating any regression style model to future climates is problematic if the relationship between predictor and response is represented in a physically unrealistic way.

Duration-specific differences in median flood estimation

We observe duration-specific differences in the partial response curves for the predictors A_{LE} (effective lake percentage) and A_P (catchment shape). These differences can be described as changes in the predictors' magnitude of effect; that is, the y axis range of the partial response curves for A_{LE} and A_P tends to be larger at shorter durations than longer durations. This means that floodGAM finds the influence of effective lake percentage and catchment shape on the median flood to be more pronounced at shorter durations.

These results indicate that the relationship between catchment descriptors and the median flood changes with duration. This means that in order to optimally model each duration, the form of the functional relationship between catchment descriptors and the median flood should be (i) adapted and (ii) re-estimated at each duration.

We examined how performance changes when these requirements are relaxed in various ways. First, to assess the performance when assuming a fixed relationship between median flood and predictors, we employed the floodGAM fitted on one duration to make predictions for another duration. We found that using the relationships established by floodGAM for one

duration to predict for another led to diminished predictive performance. Secondly, to assess performance when assuming a parametric relationship and re-estimating the model for each duration, we fit RFFA_2018—which was developed for the 24 hour data–to the 1 hour data. Once again, the performance was lower compared to the fully flexible duration-specific model, although assuming a fixed parametric form and re-estimating yielded better results than assuming an entirely fixed relationship (without re-estimating the coefficients). In the context of models that simultaneously estimate the median flood at different durations, this suggests it would be challenging to achieve optimal outcomes for every duration. In such scenarios, practitioners might need to decide on which durations reduced performance would be acceptable.

Predictor selection

Our study was focused on the question: does a data-driven model (floodGAM) detect duration-specific differences in how catchment covariates influence the median flood? If we gauge model adequacy through predictive performance, we are naturally confined to answering our question within predictor sets that work well with these sorts of data-driven models.

Identification of predictor sets that are good for data-driven models can be interesting in and of itself as it is possible that data-driven models can uncover predictor information that was previously unclear (Guyon and Elisseeff, 2003). The challenge here is that using a data-driven model for selection implies in most cases a model-based preselection, which is not guaranteed to generate a predictor set that will work within other model architectures (Maier et al., 2010). In this study, one type of data-driven model (a boosted tree ensemble with a depth of one) was used to preselect a predictor set that was then validated inside a different type of data-driven model architecture (the GAM). The selection of the predictor set by two different data-driven models suggests a certain degree of robustness. However, we do not necessarily expect the chosen predictor set in this study to give good results when used with an entirely different model architecture, e.g. a log-linear model.

This limits cross-model architecture and cross-predictor set questions. For example, we cannot say if the differences in performance between floodGAM and RFFA_2018 at the 1 hour duration are due to the fixed functional form assumed in RFFA_2018 or the differences in predictor sets, although the duration-specific differences identified within floodGAM suggest that it is advantageous to be able to adapt to different predictor-response relationships at different durations. Answering questions focusing on the duration dependence of particular catchment descriptors or predictor sets is an interesting area of future research that requires hydrology-specific knowledge reflecting a mechanistic understanding of the process at hand.

CONCLUSIONS

We develop a generalized additive modeling approach for estimation of the median annual maximum (index) flood, with a focus on detection and description of the functional relationships between the median flood and catchment descriptors at multiple durations. We employ a machine learning-based variable to flag promising predictors for expert assessment and increase the practicality of constructing generalized additive models (GAMs) for index flood estimation. We establish the adequacy of the GAM as an explainable model through predictive performance at ungauged locations, where predictive performance was measured by both predictive accuracy and reliability. The predictive performance of the GAM developed in this study (flood-

GAM) is compared to two benchmark models, the existing log-linear model for median flood estimation in Norway and a fully data-driven machine learning model (an extreme gradient boosting tree ensemble, XGBoost). We find that

- The predictive accuracy and reliability of floodGAM matches or exceeds that of the benchmark models at all durations studied.
- We observe duration-specific differences in the form of the functional relationship between the median flood and two catchment descriptors (effective lake percentage and catchment shape) within the predictor set considered in floodGAM. Ignoring these differences results in a statistically significant decline in predictive performance.

If index flood estimation at multiple durations is the goal, these results suggest that it may be difficult to obtain optimal performance on all durations when assuming a fixed or parametric form between predictors and response. Models and approaches that make these assumptions while accounting for, or extrapolating to, different durations should consider on which durations it would be acceptable to have reduced performance. Finally, in situations where predictive performance at multiple observed durations is a priority, floodGAM emerges as a promising option. The ability to auto-adapt functional relationships at multiple durations offers a potential simplification of the modeling process and could be a practical alternative to development of separate parametric forms. Furthermore, the comparative predictive performance between floodGAM and XGBoost suggests that floodGAM is adequately capturing the available relationships in the data, while also providing interpretability and accessible information on prediction uncertainty.

Data availability. The flood and hydrological data were extracted from the National Hydrological Database (Hydra II) hosted by the Norwegian Water Resources and Energy Directorate (NVE). The data used in this analysis are published at https://doi.org/10.5281/zenodo.8415076

Author contributions. DMB, KE, TLT developed the concept and TK and CYX supported the analyses. DMB performed the formal analyses, produced the figures, and wrote the first draft of the paper, which was revised by KE, TK, TLT, CYX

Competing interests. The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: Danielle Barna, Kolbjørn Engeland, Thordis Thorarinsdottir and Chong-Yu Xu report financial support was provided by Research Council of Norway.

Acknowledgements. The authors would like to thank Mads-Peter Dahl for help with data selection.

Financial support. This work was supported by the Research Council of Norway through grant nr. 302457 “Climate adjusted design values for extreme precipitation and flooding” (ClimDesign) and FRINATEK Project 274310.

References

Alexander, G. (1972). Effect of catchment area on flood magnitude. *Journal of Hydrology*, 16(3):225–240.

Alsahaf, A., Petkov, N., Shenoy, V., and Azzopardi, G. (2022). A framework for feature selection through boosting. *Expert Systems with Applications*, 187:115895.

Barna, D. M., Engeland, K., Thorarinsdottir, T. L., and Xu, C.-Y. (2023). Flexible and consistent flood–duration–frequency modeling: A bayesian approach. *Journal of Hydrology*, 620:129448.

Beldring, S., Roald, L., and Voksø, A. (2002). Avrenningskart for norge. årsmiddelverdier for avrenning 1961-1990 [map of annual runoff for norway for the period 1961-1990]. Technical Report 36.

Blöschl, G. (2013). *Runoff prediction in ungauged basins: synthesis across processes, places and scales*. Cambridge University Press.

Blöschl, G. and Sivapalan, M. (1995). Scale issues in hydrological modelling: a review. *Hydrological processes*, 9(3-4):251–290.

Breinl, K., Lun, D., Müller-Thomy, H., and Blöschl, G. (2021). Understanding the relationship between rainfall and flood probabilities through combined intensity-duration-frequency analysis. *Journal of Hydrology*, 602(March):126759.

Chebana, F., Charron, C., Ouarda, T. B., and Martel, B. (2014). Regional frequency analysis at ungauged sites with the generalized additive model. *Journal of Hydrometeorology*, 15(6):2418–2428.

Chen, T. and Guestrin, C. (2016). Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 785–794.

Chen, T., He, T., Benesty, M., Khotilovich, V., Tang, Y., Cho, H., Chen, K., Mitchell, R., Cano, I., Zhou, T., et al. (2015). Xgboost: extreme gradient boosting. *R package version 0.4-2*, 1(4):1–4.

David, V. and Davidova, T. (2014). Methodology for flood frequency estimations in small catchments. *Natural Hazards and Earth System Sciences*, 14(10):2655–2669.

Dawid, A. P. (1984). Present position and potential developments: Some personal views statistical theory the prequential approach. *Journal of the Royal Statistical Society: Series A (General)*, 147(2):278–290.

Dubos, V., Hani, I., Ouarda, T. B., and St-Hilaire, A. (2022). Short-term forecasting of spring freshet peak flow with the generalized additive model. *Journal of Hydrology*, 612:128089.

Engeland, K., Glad, P., Hamududu, B. H., Li, H., Reitan, T., and Stenius, S. M. (2020). Lokal og regional flomfrekvensanalyse [local and regional flood frequency analysis]. Technical Report 10, NVE.

Engeland, K., Schlichting, L., Randen, F., Nordtun, K. S., Reitan, T., Wang, T., Holmqvist, E., Voksø, A., and Eide, V. (2016). Utvalg og kvalitetssikring av flomdata forflomfrekvensanalyser [choice and quality control of flood data for flood frequency analyses]. Technical Report 85, NVE.

Fischer, S. and Schumann, A. H. (2021). Regionalisation of flood frequencies based on flood type-specific mixture distributions. *Journal of Hydrology X*, 13:100–107.

Friedman, J., Hastie, T., and Tibshirani, R. (2000). Additive logistic regression: a statistical view of boosting (with discussion and a rejoinder by the authors). *The Annals of Statistics*, 28(2):337–407.

Friedman, J. H. (2001). Greedy function approximation: a gradient boosting machine. *The Annals of Statistics*, 29:1189–1232.

Gaál, L., Szolgay, J., Kohnová, S., Hlavčová, K., Parajka, J., Viglione, A., Merz, R., and Blöschl, G. (2015). Dependence between flood peaks and volumes: a case study on climate and hydrological controls. *Hydrological Sciences Journal*, 60(6):968–984.

Gaál, L., Szolgay, J., Kohnová, S., Parajka, J., Merz, R., Viglione, A., and Blöschl, G. (2012). Flood timescales: Understanding the interplay of climate and catchment processes through comparative hydrology. *Water Resources Research*, 48(4).

Galelli, S. and Castelletti, A. (2013). Tree-based iterative input variable selection for hydrological modeling. *Water Resources Research*, 49(7):4295–4310.

Gneiting, T. (2011). Making and evaluating point forecasts. *Journal of the American Statistical Association*, 106(494):746–762.

Gneiting, T. and Raftery, A. E. (2007). Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association*, 102(477):359–378.

Gräler, B., Van Den Berg, M., Vandenberghe, S., Petroselli, A., Grimaldi, S., De Baets, B., and Verhoest, N. (2013). Multivariate return periods in hydrology: a critical and practical review focusing on synthetic design hydrograph estimation. *Hydrology and Earth System Sciences*, 17(4):1281–1296.

Guyon, I. and Elisseeff, A. (2003). An introduction to variable and feature selection. *Journal of Machine Learning Research*, 3(Mar):1157–1182.

Hastie, T. and Tibshirani, R. (1987). Generalized additive models: some applications. *Journal of the American Statistical Association*, 82(398):371–386.

He, S., Guo, S., Zhang, J., Liu, Z., Cui, Z., Zhang, Y., and Zheng, Y. (2022). Multi-objective operation of cascade reservoirs based on short-term ensemble streamflow prediction. *Journal of Hydrology*, 610:127936.

Hegdahl, T. J., Engeland, K., Steinsland, I., and Tallaksen, L. M. (2019). Streamflow forecast sensitivity to air temperature forecast calibration for 139 norwegian catchments. *Hydrology and Earth System Sciences*, 23(2):723–739.

Hersbach, H. (2000). Decomposition of the continuous ranked probability score for ensemble prediction systems. *Weather and Forecasting*, 15(5):559–570.

Laimighofer, J., Melcher, M., and Laaha, G. (2022a). Low-flow estimation beyond the mean–expectile loss and extreme gradient boosting for spatiotemporal low-flow prediction in austria. *Hydrology and Earth System Sciences*, 26(17):4553–4574.

Laimighofer, J., Melcher, M., and Laaha, G. (2022b). Parsimonious statistical learning models for low-flow estimation. *Hydrology and Earth System Sciences*, 26(1):129–148.

Lamontagne, J. R., Stedinger, J. R., Berenbrock, C., Veilleux, A. G., Ferris, J. C., and Knifong, D. L. (2012). Development of regional skews for selected flood durations for the central valley region, california, based on data through water year 2008. Technical Report 5130, US Geological Survey.

Lun, D., Viglione, A., Bertola, M., Komma, J., Parajka, J., Valent, P., and Blöschl, G. (2021). Characteristics and process controls of statistical flood moments in europe–a data-based analysis. *Hydrology and Earth System Sciences*, 25(10):5535–5560.

Lussana, C., Tveito, O. E., Dobler, A., and Tunheim, K. (2019). senorge_2018, daily precipitation, and temperature datasets over norway. *Earth System Science Data*, 11(4):1531–1551.

Maier, H. R., Jain, A., Dandy, G. C., and Sudheer, K. P. (2010). Methods used for the development of neural networks for the prediction of water resource variables in river systems: Current status and future directions. *Environmental Modelling & Software*, 25(8):891–909.

Marra, G. and Wood, S. N. (2011). Practical variable selection for generalized additive models. *Computational Statistics & Data Analysis*, 55(7):2372–2387.

McCuen, R. H. and Beighley, R. E. (2003). Seasonal flow frequency analysis. *Journal of Hydrology*, 279(1-4):43–56.

Merz, R. and Blöschl, G. (2009). Process controls on the statistical flood moments-a data based analysis. *Hydrological Processes: An International Journal*, 23(5):675–696.

Msilini, A., Charron, C., Ouarda, T. B., and Masselot, P. (2022). Flood frequency analysis at ungauged catchments with the gam and mars approaches in the montreal region, canada. *Canadian Water Resources Journal/Revue canadienne des ressources hydriques*, 47(2-3):111–121.

Murthy, C. S. (2002). *Water resources engineering: Principles and practice*. New Age International.

Najibi, N. and Devineni, N. (2023). Scaling of floods with geomorphologic characteristics and precipitation variability across the conterminous united states. *Water Resources Research*, 59(2).

Ni, L., Wang, D., Wu, J., Wang, Y., Tao, Y., Zhang, J., and Liu, J. (2020). Streamflow forecasting using extreme gradient boosting model coupled with gaussian mixture model. *Journal of Hydrology*, 586:124901.

Noor, F., Laz, O. U., Haddad, K., Alim, M. A., and Rahman, A. (2022). Comparison between quantile regression technique and generalised additive model for regional flood frequency analysis: A case study for victoria, australia. *Water*, 14(22):3627.

Ouarda, T. B., Cunderlik, J. M., St-Hilaire, A., Barbet, M., Bruneau, P., and Bobée, B. (2006). Data-based comparison of seasonality-based regional flood frequency methods. *Journal of Hydrology*, 330(1-2):329–339.

Pesantez, J. E., Berglund, E. Z., and Kaza, N. (2020). Smart meters data for modeling and forecasting water demand at the user-level. *Environmental Modelling & Software*, 125:104633.

Prasad, R., Deo, R. C., Li, Y., and Maraseni, T. (2017). Input selection and performance optimization of ann-based streamflow forecasts in the drought-prone murray darling basin region using iis and modwt algorithm. *Atmospheric Research*, 197:42–63.

Rahman, A., Charron, C., Ouarda, T. B., and Chebana, F. (2018). Development of regional flood frequency analysis techniques using generalized additive models for australia. *Stochastic Environmental Research and Risk Assessment*, 32:123–139.

Requena, A. I., Chebana, F., and Mediero, L. (2016). A complete procedure for multivariate index-flood model application. *Journal of Hydrology*, 535:559–580.

Robinson, J. S. and Sivapalan, M. (1997a). An investigation into the physical causes of scaling and heterogeneity of regional flood frequency. *Water Resources Research*, 33(5):1045–1059.

Robinson, J. S. and Sivapalan, M. (1997b). Temporal scales and hydrological regimes: Implications for flood frequency scaling. *Water Resources Research*, 33(12):2981–2999.

Robson, A. and Reed, D. (1999). *Flood Estimation Handbook. Vol. 3: Statistical Procedures for Flood Frequency Estimation*. Institute of Hydrology.

Saloranta, T. (2014). New version (v.1.1.1) of the seNorge snow model and snow maps for Norway. Technical Report 6, Norges Vassdrags og Energidirektorat (NVE).

Stedinger, J. R. (1980). Fitting log normal distributions to hydrologic data. *Water Resources Research*, 16(3):481–490.

Stein, L., Clark, M. P., Knoben, W. J., Pianosi, F., and Woods, R. A. (2021). How do climate and catchment attributes influence flood generating processes? a large-sample study for 671 catchments across the contiguous usa. *Water Resources Research*, 57(4).

Tarasova, L., Basso, S., Poncelet, C., and Merz, R. (2018). Exploring controls on rainfall-runoff events: regional patterns and spatial controls of event characteristics in germany. *Water Resources Research*, 54(10):7688–7710.

The Norwegian Water and Energy Directorate (2010). Elvis elvenett - geonorge datasett [elvis river networks - geonorge dataset].

Thorarinsdottir, T. L., Sillmann, J., Haugen, M., Gissibl, N., and Sandstad, M. (2020). Evaluation of cmip5 and cmip6 simulations of historical surface air temperature extremes using proper evaluation methods. *Environmental Research Letters*, 15(12):124041.

Tsonis, A. A., Elsner, J. B., Gupta, V. K., Troutman, B. M., and Dawdy, D. R. (2007). Towards a nonlinear geophysical theory of floods in river networks: an overview of 20 years of progress. *Nonlinear Dynamics in Geosciences*, pages 121–151.

Viglione, A. and Blöschl, G. (2009). On the role of storm duration in the mapping of rainfall to flood return periods. *Hydrology and Earth System Sciences*, 13(2):205–216.

Viglione, A., Chirico, G. B., Komma, J., Woods, R., Borga, M., and Blöschl, G. (2010). Quantifying space-time dynamics of flood event types. *Journal of Hydrology*, 394(1-2):213–229.

Vormoor, K., Lawrence, D., Schlichting, L., Wilson, D., and Wong, W. K. (2016). Evidence for changes in the magnitude and frequency of observed rainfall vs. snowmelt driven floods in norway. *Journal of Hydrology*, 538:33–48.

Wood, S. (2017). *Generalized Additive Models: An Introduction with R*. Chapman and Hall/CRC, 2 edition.

Zoccatelli, D., Borga, M., Viglione, A., Chirico, G. B., and Blöschl, G. (2011). Spatial moments of catchment rainfall: rainfall spatial organisation, basin morphology, and flood response. *Hydrology and Earth System Sciences*, 15(12):3767–3783.

Zounemat-Kermani, M., Batelaan, O., Fadaee, M., and Hinkelmann, R. (2021). Ensemble machine learning paradigms in hydrology: A review. *Journal of Hydrology*, 598:126266.

Appendix: Computation of minimizing quantity for relative error and absolute percent error

The optimal predictor for the relative error is the functional $\text{med}^{(1)}(F)$, defined in Gneiting (2011), which is the median of the distribution with density proportional to $xf(x)$. Here $f(x)$ is the probability density function for the log normal distribution; that is:

$$f(x) = \frac{1}{x\sigma\sqrt{2\pi}} \exp\left(-\frac{(\ln x - \mu)^2}{2\sigma^2}\right) \qquad x > 0 \tag{1}$$

and the density proportional to $xf(x)$ is given by $g(x) = 1/A \cdot xf(x)$, where A is a normalizing constant such that $g(x)$ is a density. Denote by G the distribution with density g. We approximate the median of G by numerically integrating $g(x)$ from 0 to m in R and conducting a grid search for the closest value of m such that $g(m) \approxeq 0.5$ on a grid with spacing 0.01. The optimal predictor for the absolute percent error is the functional $\text{med}^{(-1)}(F)$–that is, the median of the distribution with density proportional to $f(x)/x$–and is found with the same approximation method.

Appendix: Hyperparameter tuning for XGBoost models within the IIS algorithm

XGBoost is used twice in this study: once as the underlying model in the Iterative Input Selection (IIS) algorithm and once as a predictive performance benchmark in the results section. The two applications are very different and require different hyperparameters. In both cases, suitable hyperparameters were chosen by grid-search and cross validation. Here we report the hyperparameter optimization set up used in the XGBoost models for the IIS algorithm

We used squared error loss as the objective function and tuned the following hyperparameters on the indicated ranges: the percentage of observations subsampled at each boosting step (0.1-1); the minimum number of instances needed in each node (1-7); and the shrinkage parameter η (0.01-0.1). The number of boosting iterations was evaluated up to a maximum number of 999 iterations. For the XGBoost models used within the IIS algorithm, tree depth was fixed at 1. Hyperparameter tuning was conducted on a grid search within a 10-fold cross-validation scheme using all possible parameter combinations and an early

stopping criterion for the number of boosting iterations, where the algorithm stopped after 25 rounds without improvement in the error rate as determined by a chosen evaluation metric. The ranges of the hyperparameters were chosen based on experience with the data set and recommended XGBoost practices. Hyperparameters were optimized separately for the 1 hour and 24 hour durations. The evaluation metric used in hyperparameter tuning cross validation is the MAE.

Appendix: Details of the machine learning based pre-selection step

Use of XGBoost within the modular structure of the IIS algorithm was first proposed by Alsahaf et al. (2022). The primary benefit to this is that use of a boosted tree ensemble—rather than a bagged tree ensemble such as the Extra-Trees routine originally proposed in Galelli and Castelletti (2013)—solves the issue of significance splitting in the input ranking algorithm. Significance splitting is when the importance scores of two or more redundant variables are split evenly. This can occur when the tree ensemble is subsampled or bootstrapped, as in bagging and random forest. The algorithm in Galelli and Castelletti (2013) accounts for this by including a secondary evaluation step of the variable ranking to reduce the impact of significance splitting. However, the success of this secondary step is reliant on hyperparameter choice (Galelli and Castelletti, 2013). The use of a boosted tree ensemble inherently solves this issue. Selecting XGBoost as the boosted tree ensemble is a natural choice: it has established use in hydrology (Zounemat-Kermani et al., 2021), is computationally efficient, is available as a package in both R and Python, and has a large and active user base.

Within the IIS algorithm, we use the additive gain as the importance score in the input ranking algorithm. As part of the model-fitting process XGBoost uses a scoring function that takes into account the improvement in the objective function (in this case, mean squared error) resulting from the inclusion of each variable. The additive gain of a variable is the sum of its gain across all boosting rounds. For details, see Chen et al. (2015). For a more robust approach, we adopt the method proposed in Laimighofer et al. (2022b), where the initial variable ranking is averaged over 25 bootstrap samples. Then the gain of each variable for the final variable ranking is the ratio of the individual additive gain to the total gain over all variables. The hyperparameters are the same both for the input ranking and the model used to test predictive performance of an additional variable.

The automatic stopping condition in IIS requires choice of both a suitable distance metric for measuring predictive accuracy of the chosen variable set and a threshold value above which a change in predictive accuracy between the proposed sets is considered insignificant. We used mean absolute error (MAE) as the distance metric. We set the threshold value to 0.1, meaning we stop selecting new variable sets when the new set results in, on average, a less than 1 l/s/km2 improvement in median flood prediction. The evaluation of the distance metric takes place across a k-fold cross validation approach to increase robustness. The dataset is divided into k mutually exclusive subsets of equal size, and the predictive model is fit k times. In each iteration, the model is validated on one of the k folds and calibrated using the other $k-1$ folds. The predicted accuracy associated with adding a particular feature is estimated as the average value of the chosen metric over the k validations. We used 10-fold validation for our data set.

Table 1. Regions are mid-south-west Norway and eastern Norway + Finmark. The "region" model is fit on the subregions, and evaluation metrics are calculated from all stations included in analysis. There were no statistically significant differences between the two models on either duration.

Duration	Name	Evaluation metric				
		RMSE [l/s/km2]	CRPS [l/s/km2]	MAE [l/s/km2]	MRE [%]	MAPE [%]
1 hour	floodGAM	122.2	61.2	84.9	20.4	20.5
	floodGAM, regions	120.5	59.1	82.8	19.6	19.2
24 hours	floodGAM	84.1	42.7	59.5	17.0	17.5
	floodGAM, regions	87.4	43.8	61.6	16.7	17.8

The main computational burden of the IIS algorithm is in this repeated model fitting required for computation of the distance metric in the k-fold cross validation: if m potential variables are evaluated at each step, $m * k$ models must be fit. Thus the choice of the number of top-ranked variables to evaluate at each step is important for model performance. While the variable ranking at each step is always computed over the entire variable set, the search space (i.e. the number of variables individually evaluated for predictive performance) can be reduced to a user-specified number of variables. We used the top 15 variables.

In this study the IIS algorithm is used within a resampling scheme. We split our data set into ten non-overlapping folds and repeatedly apply the IIS algorithm while withholding one of the folds at a time. A visual explanation of the IIS algorithm and this resampling scheme is found in Fig. 1.

Full grid output

Appendix: Supplementary figures for model evaluation metrics

Appendix: Regional assessment

The predictive performance of floodGAM was not significantly improved by splitting the area of study (Norway) into hydrologically homogeneous regions and fitting floodGAM within the regions. Two regions were used: mid-, south and west Norway (region 1) and east Norway and Finmark (region 2); see Fig. 8. The regions are those defined in Hegdahl et al. (2019). Within each region, we ran a 10-fold cross validation; predictive performance metrics were calculated between regional model predictions and the hold out data for the 1 and 24 hour durations. The performance metrics were then summarized across the whole of Norway so they could be compared to the metrics from floodGAM fit to the entire country. Note that this means the hold-out data between floodGAM and “floodGAM, regions” is not identical; however, given the similarities between the metrics it is unlikely this variation in the cross-validation is very influential. Table 1 reports the evaluation metrics from both floodGAM fit to the entire country and floodGAM fit within regions. There were no statistically significant differences between the reported evaluation metrics at either duration.

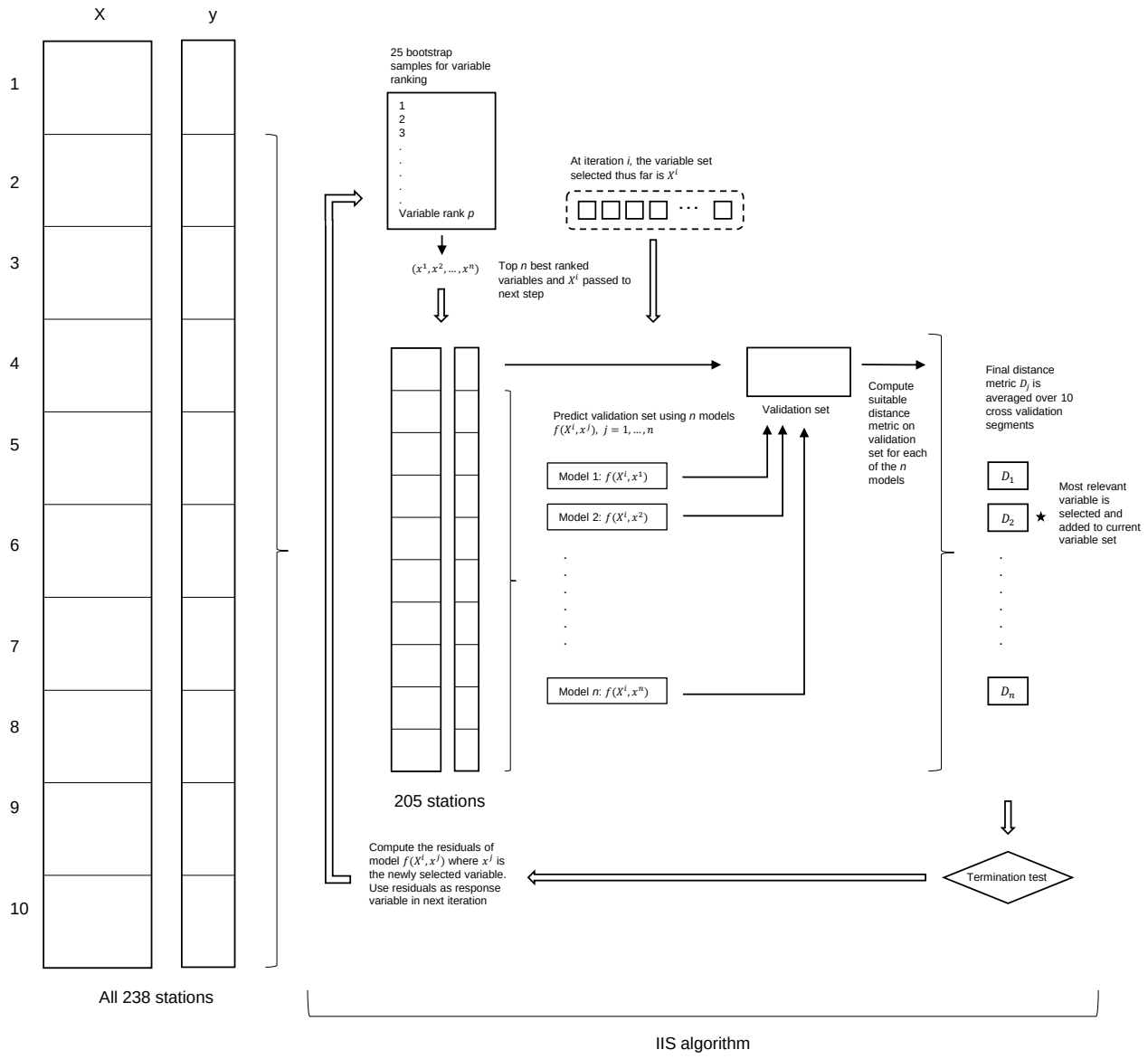

Figure 1. Visual depiction of the variable pre-selection scheme, showing the IIS algorithm and the resampling method.

Appendix: Seasonal variations in hydro-climatic predictors

To further investigate the relationships found by floodGAM, we compute seasonal maxima for two seasons: a summer season from April-July and a winter season from August-March for the 1 and 24 hour durations. The full model—all seven predic-

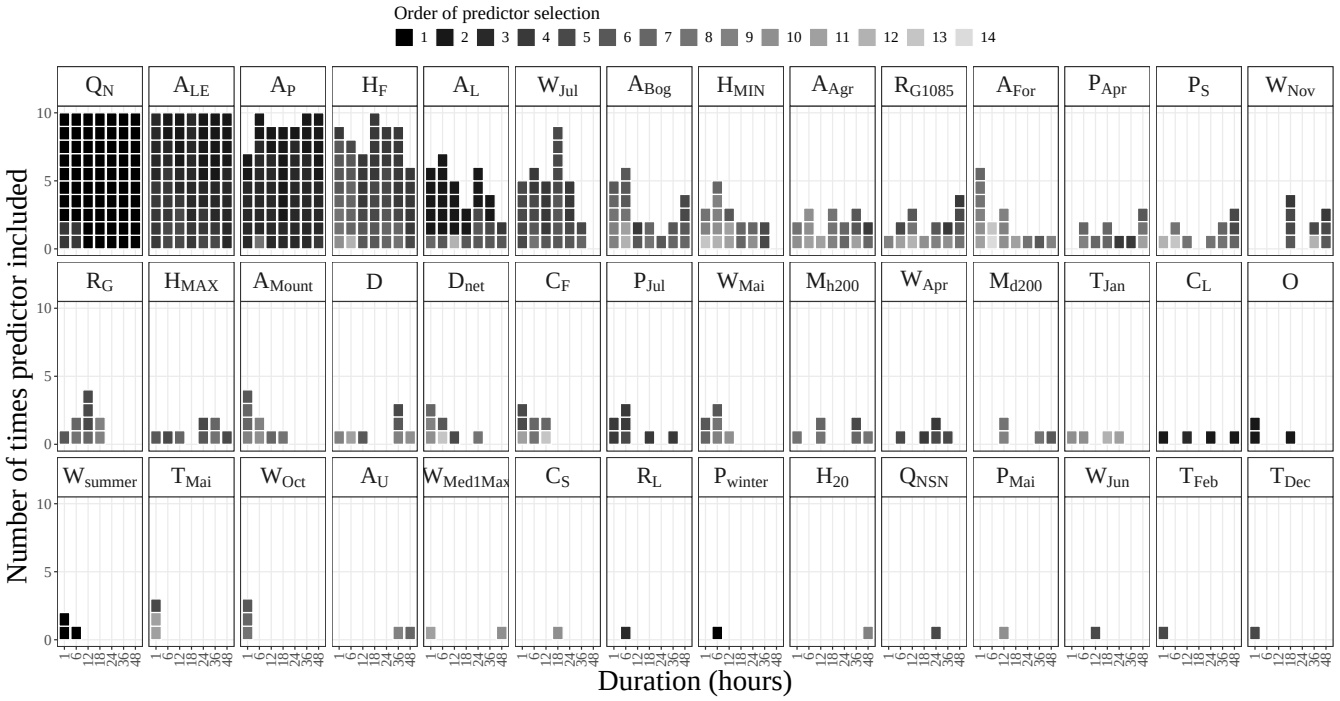

Figure 2. Full results from variable pre-selection. The vertical axis represents the number of times a variable was chosen. The horizontal axis indicates the duration that generated the covariate set. The color indicates the order of variable selection within the IIS algorithm. Variables selected first tend to be those that are most informative.

tors–is fit on both seasons, and the select argument of the gam() function is set to 'True' such that any predictors found to be irrelevant can be shrunk out of the seasonal models. This serves as a check on the significance of the two climate characteristics (W_{Apr} and P_{Sep}) for the seasonal maxima since April is excluded from the winter months, and September is excluded from the summer months. Figure 9 shows the estimated smooth components for the seasonal maxima along with the associated 95 % estimation uncertainty intervals. The estimated smooth components for the annual maxima are underlaid as dashed grey lines. The location of data points for each predictor is indicated along the x axis, and y axis ranges are duplicated from Fig. 7.

Figure 9 shows that all four geographical catchment descriptors and one of the hydro-climatic cdescriptors–Q_N–have a consistent shape across durations and seasons. However, splitting on seasons changes the shape of the smooth components for the other two hydro-climatic characteristics, W_{Apr} and P_{Sep}. For the spring season, P_{Sep} is shrunk out of the model entirely whereas the relationship between W_{Apr} and median spring floods is decreasing.

The autumn/winter seasonal maxima, however, find both W_{Apr} and P_{Sep} to be significant. The shapes of the relationships between the catchment descriptors for the annual model is replicated in the autumn/winter season, except for P_{Sep} where a slightly concave relationship is seen.

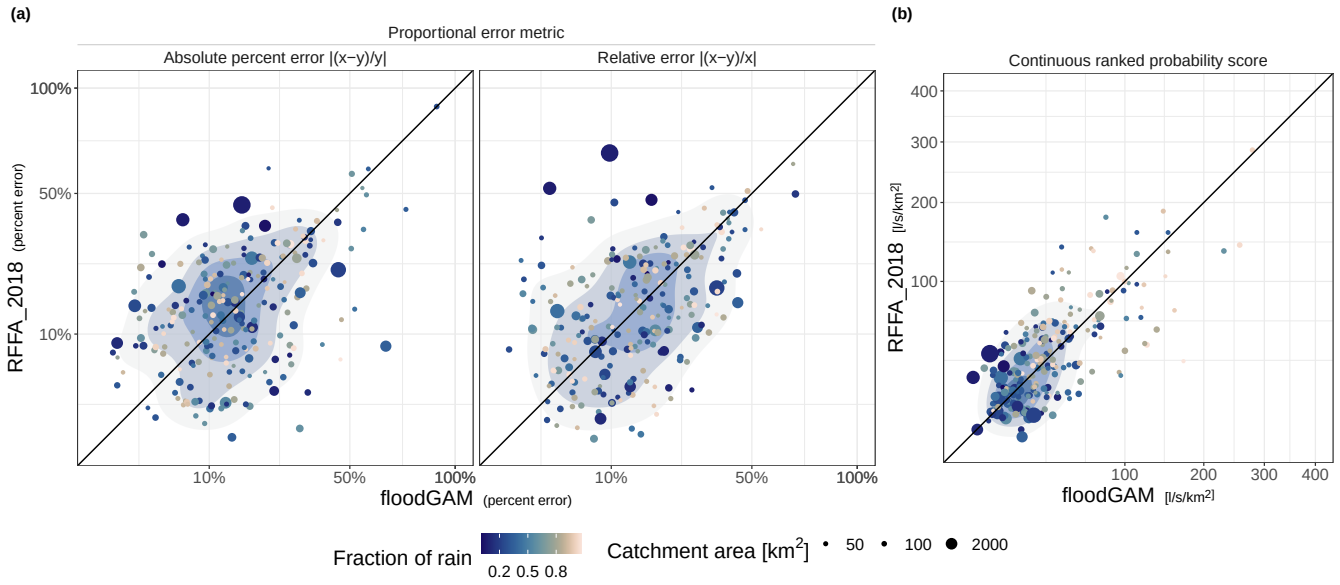

Figure 3. Model to model comparison on absolute percent error, relative error, and the continuous ranked probability score for RFFA_2018 and floodGAM on the 24 hour duration. In the panel headers, x represents the predicted value and y the observed value. Points falling above the diagonal line indicate stations where RFFA_2018 performed worse than floodGAM. Points falling below the diagonal line indicate stations where floodGAM performed worse than RFFA_2018. The 2D kernel density estimation of point density is underlaid to aid visual interpretation. Point size shows catchment area, point color indicates the fraction of rain contribution to flood.

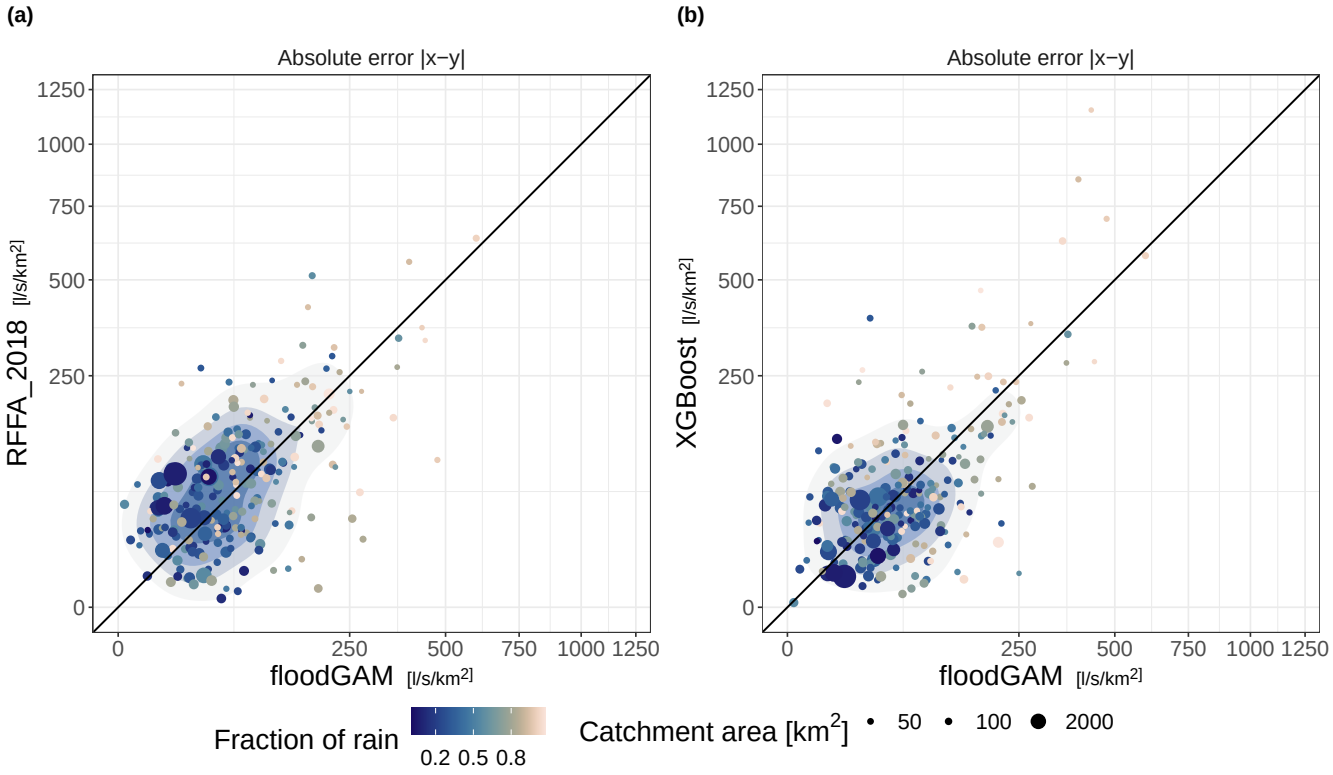

Figure 4. Model to model comparison on absolute error for RFFA_2018 vs floodGAM and XGBoost vs floodGAM on the 1 hour duration. In the panel headers, x represents the predicted value and y the observed value. Points falling above the diagonal line indicate stations where the comparative model (RFFA_2018 or XGBoost) performed worse than floodGAM, and vice versa for points falling below the diagonal line. The 2D kernel density estimation of point density is underlaid to aid visual interpretation. Point size shows catchment area, point color indicates the fraction of rain contribution to flood.

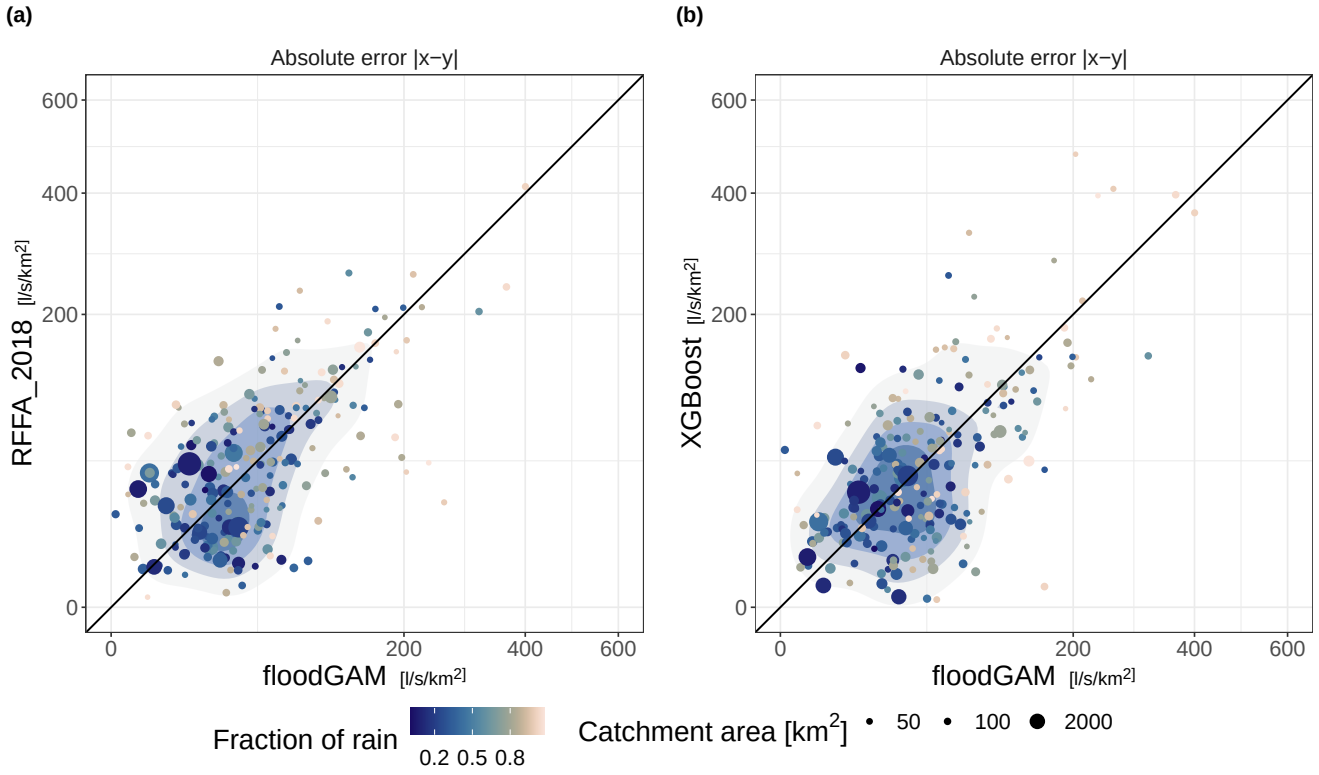

Figure 5. Model to model comparison on absolute error for RFFA_2018 vs floodGAM and XGBoost vs floodGAM on the 24 hour duration. In the panel headers, x represents the predicted value and y the observed value. Points falling above the diagonal line indicate stations where the comparative model (RFFA_2018 or XGBoost) performed worse than floodGAM, and vice versa for points falling below the diagonal line. The 2D kernel density estimation of point density is underlaid to aid visual interpretation. Point size shows catchment area, point color indicates the fraction of rain contribution to flood.

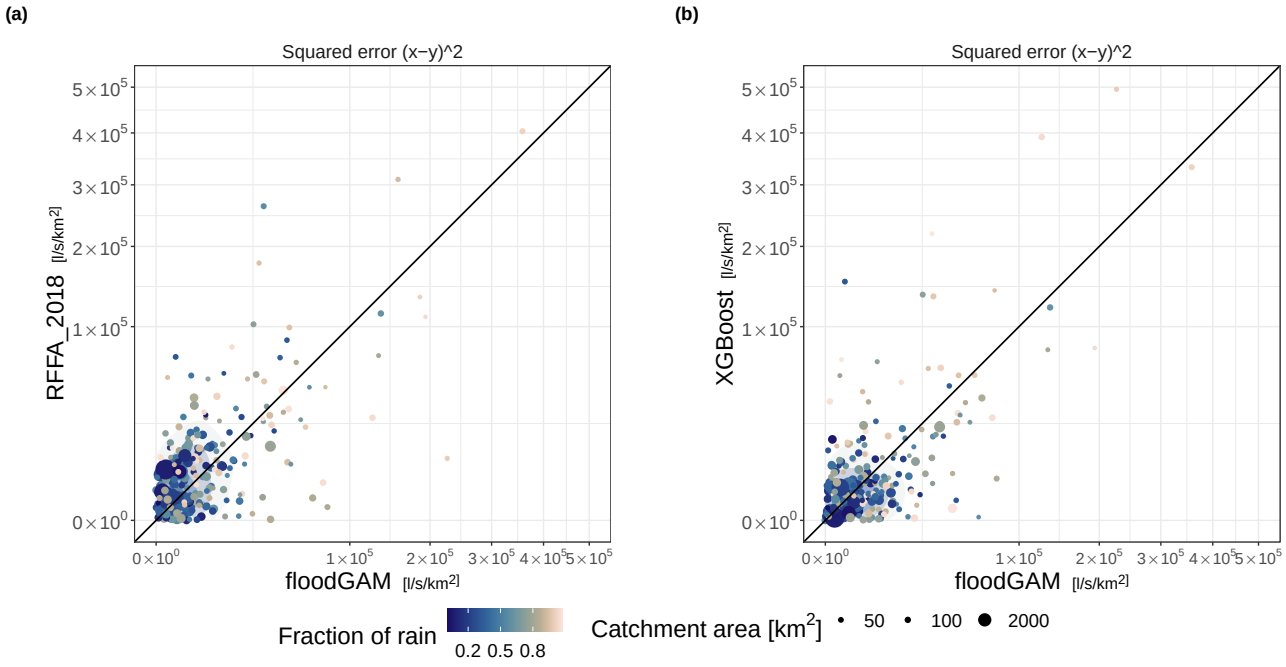

Figure 6. Model to model comparison on squared error for RFFA_2018 vs floodGAM and XGBoost vs floodGAM on the 1 hour duration. In the panel headers, x represents the predicted value and y the observed value. Points falling above the diagonal line indicate stations where the comparative model (RFFA_2018 or XGBoost) performed worse than floodGAM, and vice versa for points falling below the diagonal line. The 2D kernel density estimation of point density is underlaid to aid visual interpretation. Point size shows catchment area, point color indicates the fraction of rain contribution to flood.

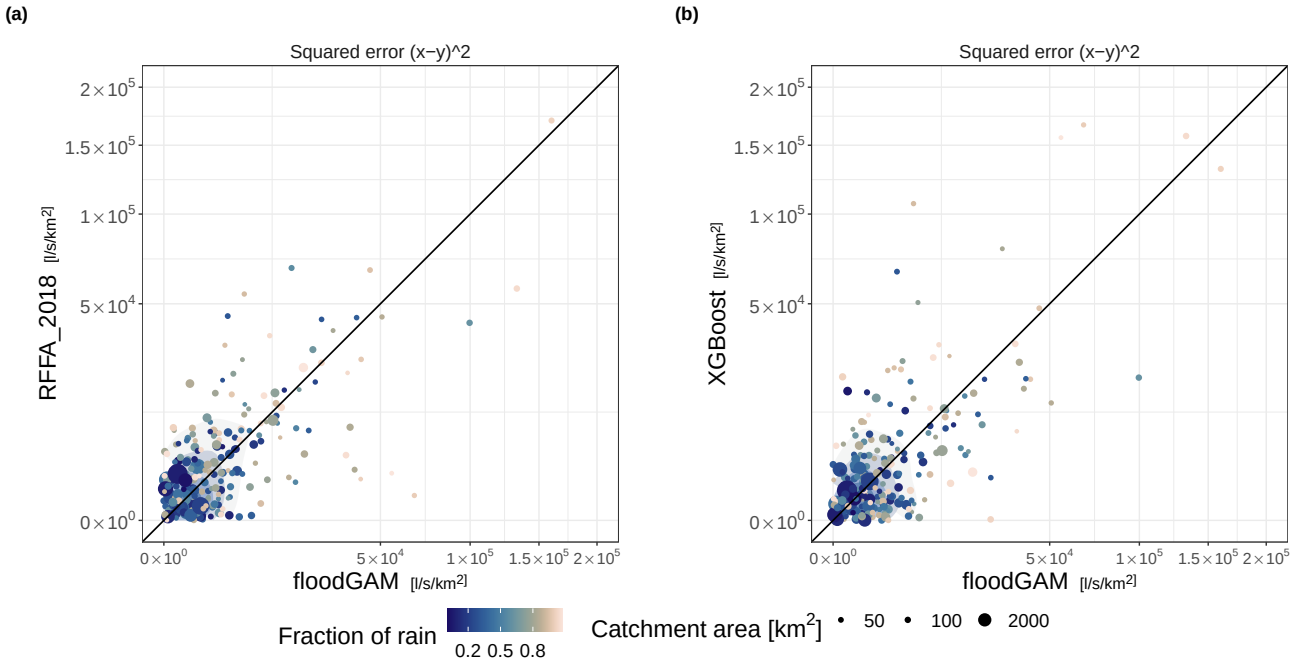

Figure 7. Model to model comparison on squared error for RFFA_2018 vs floodGAM and XGBoost vs floodGAM on the 24 hour duration. In the panel headers, x represents the predicted value and y the observed value. Points falling above the diagonal line indicate stations where the comparative model (RFFA_2018 or XGBoost) performed worse than floodGAM, and vice versa for points falling below the diagonal line. The 2D kernel density estimation of point density is underlaid to aid visual interpretation. Point size shows catchment area, point color indicates the fraction of rain contribution to flood.

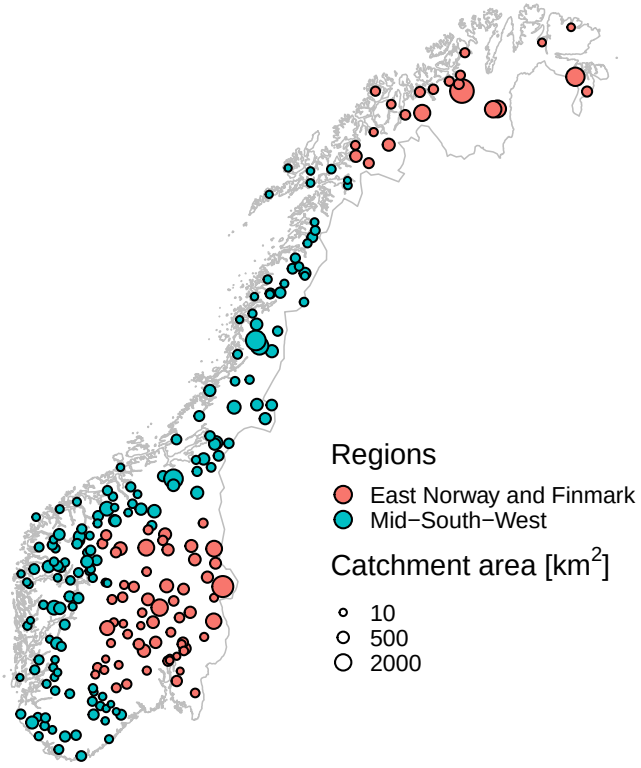

Figure 8. Regional groupings for the 232 stations used in this study. Regions are mid-, south and west Norway and east Norway and Finmark; these regions are defined in Hegdahl et al. (2019).

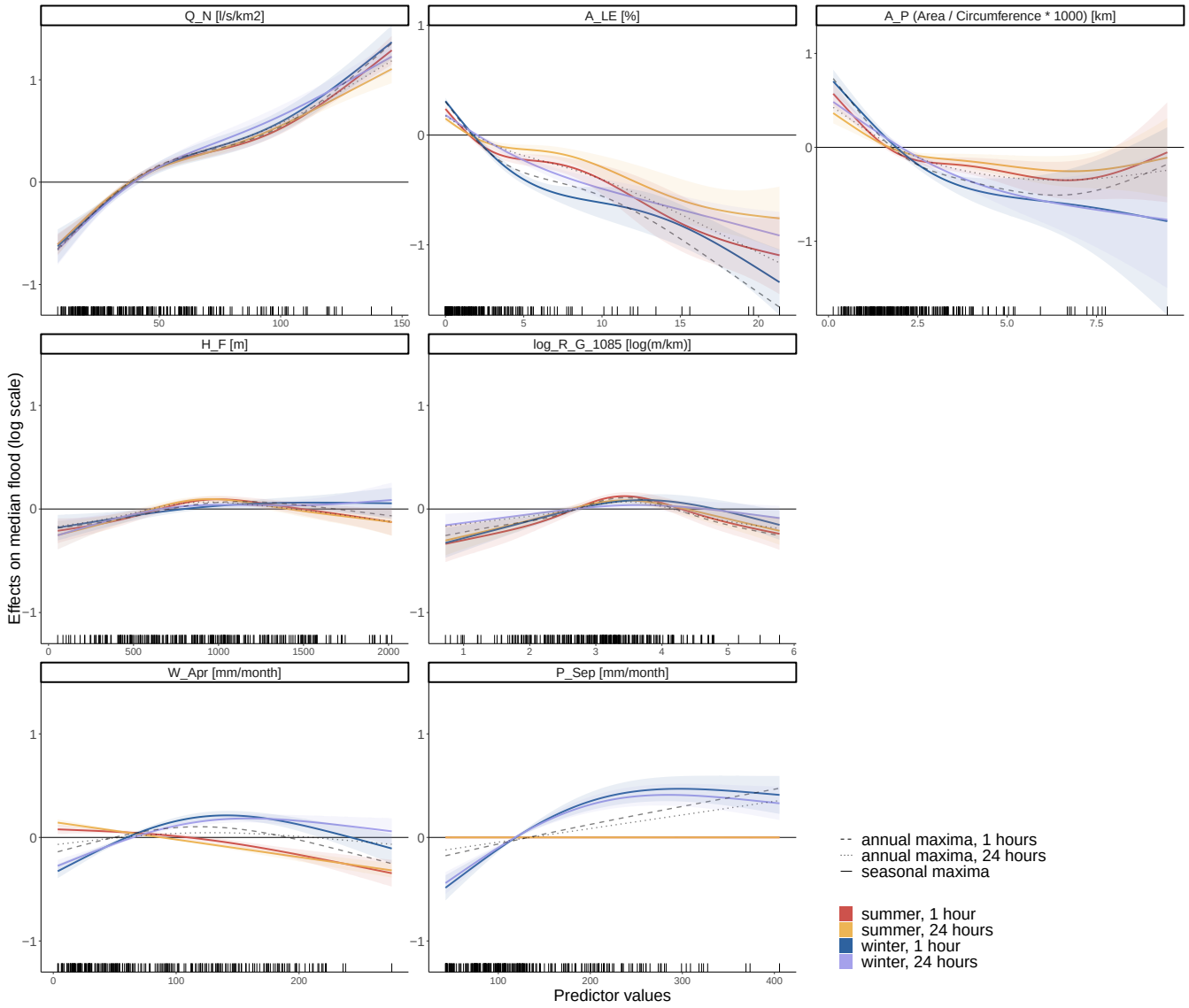

Figure 9. Partial response curves by season and duration. The summer season is April-July. The winter season is August-March. 95 % estimation uncertainty intervals for the seasonal smooth components are shown in shading, and smooth components from the annual maxima model are underlaid as dashed or dotted lines. A flat effect means a predictor was selected out of the model by shrinkage. P_{Sep} is selected out of the summer season. The smooth components are shown on the link scale, and units for predictors are shown in panel titles. Location of data points for each predictor are shown as tick marks on the x axis. Y axis ranges span the same magnitude for each panel.