What is RFFA – regression for RFFA

Mulitple flood durations and why we need multiple estimates

How to incorporate duration into design values: event-based vs aggregation

Regression models at different durations require special considerations

There is a gap – we build models independently for each duration. We must be able to compare to existing models

What is a GAM – advantages, description

Where have GAMs been used before

Why predictor selection is difficult in GAMs…shrinkage based methods

Need to select a small, uncorrelated set of predictors

Why we describe our predictor selection method here

Objectives and research questions

Flow-duration curves

Moving window analyses

Copulas and multivariate analyses

Synthetic design hydrographs

A common challenge in engineering design for retention-based systems is the need for duration-specific design values at ungauged locations. These applications often require frequency estimates of average streamflow for pre-determined durations. Moving-window analyses (alternatively: n-day flood) enable frequency analysis of the period with the highest average streamflow over n consecutive hours; see, for example . Frequency analyses are usually performed separately for each pre-determined duration, although some approaches (Barna, Cunderlik, Ourada, Javelle) attempt to scale estimates across durations by fitting an “average” distribution under strict assumptions.

In the typical case, where frequency analyses are performed separately, regression equations are typically built individually for each duration to regionalize the estimates…(or built jointly, re-estimated for each duration?)

This introduces some assumptions…

Our focus is the structure of the regionalization

*Durations at ungauged locations*

Developing regional regression equations for duration-specific methods

Nonlinear and data-driven approaches

Thorough (practically motivated) reliability and predictive performance assessment

the period of n consecutive hours with the highest average streamflow, where the streamflow is sometimes averaged over multiple flood events.

“Flood-duration flows are reported in dimensions of volume per time and units of cubic feet per second; to convert to total volume the flow rate is simply multiplied by the length of the duration interval considered.»

**Random thoughts:**

We average discharge across duration (average flowrate)

consistency

Hydrologically: The processes that produce a high average flowrate over 1 hour may be different than the processes that produce a high average flowrate over 24 hours. These differences may be so fundamentally different that we may have to adapt the functional form of the relationship, not just the model coefficients. Practically: a model would have to adapt to different relationships at different durations, or we risk imposing an artificial mathematical relationship on a changing hydrological process.

Primary: detection and description of func relation.

Recent developments leave us well-positioned to develop more accurate and more reliable regression models. Data-driven approaches. And examine potential watershed processes governing different durations.

A challenge when working with some of the newer statistical approaches is available performance metrics

Practically-oriented

We did not find evidence for different covariates at different durations, just different functional relationships…maybe because we use annual maxima.

A common challenge in engineering design for retention-based applications is the need for duration-specific design values at ungauged locations. Since these applications focus on total storage capacity, they often require frequency estimates of average streamflow over pre-determined durations. Moving-window analyses (alternatively termed sustained flood flow, N-day flood or flood-duration-frequency analyses) enable frequency analysis of the period with the highest average streamflow over n consecutive hours; see, for example . Frequency analysis is typically performed separately for each pre-determined duration. Some approaches (e.g., Barna, Cunderlik, Ourada, Javelle) attempt to scale estimates across durations by fitting an 'average' distribution. However, this scaling relies on strict assumptions that are sometimes unsupported by empirical analysis, especially when applied over broad regions (cite us and Kennedy). Separate frequency analyses for each duration are the more common approach in regional applications and are the focus of this paper.

To extend these estimates to ungauged locations, regression models are typically used. Moving-window analyses often model a wide range of durations (e.g., Lamontagne models durations from 1 to 30 days) by developing a single best parametric regression model and re-estimating the coefficients for each duration (Lamontagne, Lind, Kennedy). Relying on a single parametric form is practical and is theorized to help maintain consistency across durations (Lind).

However, this approach assumes that the same parametric form applies to all durations. This does not fully account for the fact that different processes drive high average streamflow over short vs long durations. For example, high average streamflow over short durations is often driven by intense rainfall events, while longer durations are influenced by processes like sustained snowmelt. These processes, and their interactions with catchment descriptors used in regression modeling, may be so fundamentally different that simply re-estimating coefficients may not capture the variations in relationships across durations.

Despite this, the assumption of a single parametric form is rarely challenged due to the complexities of developing and testing regional regression models on many durations. Instead, practitioners sometimes limit event types in moving-window analyses (Lamontagne) and/or allow for differing amounts of cross-correlation between stations when re-estimating regression coefficients for each duration (lamontagne, kennedy). The latter typically requires a two-step estimation procedure as high cross-correlations can occur between gauging stations in moving-window analysis, particularly at longer durations (Kennedy).

This paper investigates whether use of a single parametric form limits the predictive performance of a regional regression model applied to multiple durations. We examine whether there are statistically significant differences in consistency, predictive accuracy, reliability, and fitted relationships across different durations and different types of regression models.

Of course, establishing a need for different models at different durations is of limited usefulness if it is not practical to build the models. Developing a regional regression model is a substantial effort. Hydrologic datasets of catchment and climate characteristics often contain a large number of potential—and potentially collinear—predictors, making predictor selection a challenge. Improving model fit often involves selecting appropriate polynomial terms and transformations of the chosen predictors, which can be a labor-intensive process. And models used to produce design values must be tested on a wide variety of performance metrics that assess predictive accuracy, reliability, and uncertainty at ungauged locations.

Therefore, implementation potential played a large role when considering what types of regression models to assess. We considered several options, all of which have established use in regional hydrologic studies.

Fully parametric regression models, like linear or log-linear models, are the current state of practice in moving window analyses and are well-established in regional flood frequency analysis in general (cite). These models are straightforward to interpret and assess but require significant work to develop an appropriate parametric form. Additionally, they rely on careful selection of uncorrelated predictors, a major part of the analysis. In contrast, fully data-driven (non-parametric) models, such as the machine learning models of (e.g., Laimgihofer, Ni, [cite]), are highly adaptable, can handle large, collinear predictor sets, and do not need a defined parametric form. They are powerful methods when predictive accuracy is the only concern. But these models generally lack the distributional assumptions needed for uncertainty and reliability assessments. A middle ground can be found in *semi-parametric* regressionmodels, such as generalized additive models (GAMs) (cite cite cite), which model predictor-response relationships non-parametrically while still specifying the response distribution, enabling full statistical assessments of accuracy, reliability, and uncertainty.

We build a GAM for []. Some moving window analysis predict parameter of EV dist, others predict flood quantiles directly. For this foundational study, we focus on predicting the median annual maximum flood, which can also be location parameter. And an index flood. Area of study is Norway. We compare the GAM to existing index flood model for Norwegian catchments. On durations []. Short durations relevant for small hydropower (lots of that in Norway), long for larger reservoirs. Focus on annual max relevant to flood guidelines in Norway.

At each duration, we assess out-of-sample predictive accuracy, reliability, and consistency for the two models when estimates are extended to ungauged locations. Focus on median means predictive accuracy can be compared directly to observed data and a wide variety of performance metrics can be used.

Just as with the existing model, we specify the model once and re-estimate at each duration. However, unlike the linear, when we restimate the GAM, the underlying form is adapting. This allows for a data-driven comparison and visualization of the fitted relationships, and estimation uncertainty, at each duration.

In our context ( design values ) reliability and uncertainty highly relevant. Limit our ability to, cannot use ml. however, ML powerful and popular method. For the sake of completeness, we include it in our analysis to the extent possible. Predictive accuracy benchmark. Other researchers have also had success in using it as a method to flag potential predictors. We also consider this.

Research Qs.

Part of what distinguishes the GAM from the log-linear model is the flexible, data-driven nature of the response relationship. So it is useful to have a predictive accuracy comparison point from a fully data-driven model. We include these comparison point to the extent possible.

We build a GAM. On these durations.

How does this help us achieve our goals? – auto-adapt regression relationship at each duration and compare to the parametric model with the re-estimated coefficients.

Just like the parametric model, we only need to define it once, but when we re-estimate the GAM on each duration the model is adapting the underlying relationship (non-parametric/data-driven modeling of the predictor-response relationship).

We develop a GAM for regional median flood prediction at multiple durations. Some moving window analyses use the regression model to predict a distribution parameter for frequency analysis (lamontagne, lind), while others predict flood quantiles directly (kennedy). In this investigative study, we focus on predicting the median annual maximum flood (i.e. index flood; dalyrimple) which can serve as both a quantile and, if the reparametrization of castro-camilla is used, the location parameter of the generalized extreme value distribution. Our study area is Norway, where we compare the GAM to an existing index flood model for Norwegian catchments. The models are compared on a set of durations ranging from 1 hour to 30 days. Shorter durations are relevant to Norway’s many small hydropower plants, while longer durations apply to larger reservoirs. The focus on annual maxima aligns with Norwegian flood guidelines.

In the context of design values used operationally, reliability and uncertainty analyses are highly relevant. Limit our ability to, cannot use machine learning. however, ML powerful and popular method. For the sake of completeness, we include it in our analysis to the extent possible. Predictive accuracy benchmark. Other researchers have also had success in using it as a method to flag potential predictors. We also consider this.

Reliability and uncertainty analyses are essential for operational design values, limiting our reliance on machine learning (ML) alone.

In our context (design values for retention-based applications) reliability and uncertainty analyses are highly relevant. The operational potential limits our reliance on machine learning. However, machine learning models are powerful and popular methods. For the sake of completeness, we include machine learning in our analysis to the extent possible.

GAM also relies on small set of uncorrelated predictors. Like parametric models. But, has access to some additional, powerful selection tools (shrinkage).

Need to explain here role of XGBoost. Predictor selection, for completeness sake also included in predictive capacity to extent possible.

Flagging promising predictors – decrease burden to practitioners

Look at durations from 1 hour to 720 hours (30 days). Hourly durations are relevant for smaller hydropower stations and stream inflow to smaller reservoirs, while longer durations (24 hours to multiple days) are relevant for large reservoirs.

Look at a single data-rich quantile (the median).

Two benchmark models to establish predictive performance

Research Qs

Annual maxima

Keep predictors the same across durations

Implement and test models that could provide the basis for regional frequency models. So models that are interpretable and can be compared to existing models and assessed on variety of performance metrics

while the relationship between event and duration and duration is not one to one, as moving windows average sometime over multiple durations, generally the processes that drive high average streamflow over 1 day are different than the processes driving high average streamflow over 30 days.

Is it acknowledged problem? Lam restrict domain of event type. Kennedy restricted to spatial regions.

Either assume that a single parametric form is sufficient, or restrict event domain

This is a very limiting assumption, but practical given that there is a gap in research…and the difficulties linking moving window analyses to event-based processes.

Given the gap in research, practitioners are left with the options of either restricting the event domain or proceeding with the (very stiff) assumption that a single parametric form describes all durations.

This paper investigates the underlying assumption…

The other piece of the puzzle is that it not super practical to build a million regression models. So even if we found differences, would it matter in practice? We want to present a foundation of a result driven approach…that can be compared to existing models on a range of performance metrics…because of this we pick a GAM

Do this for a single data-rich quantile, median. Index flood.

Wide variety of durations usually modeled (1-60 days for Columbia river basin; shorter durations may be relevant for smaller hydropower stations)

* Not the most useful (is of limited utility) to establish that a more complicated relationship exists if we cannot practically model it in a framework that fits operational design value estimates
* so, need a model framework that allows for full comparison and assessment

data-driven (non and semi parametric models). GAMs. Xgboost. How GAMS offer a simplification of the model process. Predictor selection

how to introduce data-driven models. (lines 59-60)

put context on our model vs the WLS and GLS approaches

Lamgihofer fits a cross-correlation model for each duration.

What is done. “a set of regression models was developed to estimate regional N-day duration skew coefficients”

Choosing a final model often involves measuring different aspects of predicitve

And, to be useful in practice, models must be tested on a wide variety of performance metrics that assess properties like predictive accuracy, reliability, and uncertainty at ungauged locations.

And, to be useful in practice, models must be tested on a wide variety of performance metrics that assess properties like predictive accuracy, reliability, and uncertainty at ungauged locations.

Why is this a burden to practitioners? Always challenging to select covariates and model for structure (general RFFA problem). So, although there is some evidence we shouldn’t be using the same model on all durations, building and analyzing these models is time consuming and it’s generally not practical to construct a separate one for each duration.

But this assumption is not investigated. An opportunity for improvement

Say something about the implications of this:

The processes that produce a high average flowrate over 1 hour may be different than the processes that produce a high average flowrate over 24 hours. These differences may be so fundamentally different that we may have to adapt the functional form of the relationship, not just the model coefficients (do we have any evidence here?).

We examine whether there are statistically significant differences in predictive accuracy, reliability, and fitted relationships across different durations and different types of regression models.

Extremely difficult to pick parameters, test these models, and this is not discounted. Identifying these differences is no good unless they can be practically modeled.

Investigates the assumption and the practicality of doing anything different.

Just re-estimating the coefficients would hide these differences, if they exist.

But “ the increase in model complexity cannot be justified by a sufficient increase in model precision”

Given that there is evidence for this, and given that recent years have seen the rise of other options to enforce consistency in moving-window analyses (Thea). And rise of data-driven models means there are practical options to side step tedious construction of parametric regression models.

Investigate the underlying assumption that differences in n-hour floods are sufficiently explained by re-estimating the regression coefficients.

We examine whether there are statistically significant differences in predictive accuracy, reliability, and fitted relationships across different durations and different types of regression models.

And: if there is a more complex model, are there sufficient increases in precision to be able to justify?

**Decide on dataset-hyfin?**

**USGS reports**

* **Address duration consistency?**

For two years, the data section at NVE has been working to create a database of high-quality “fine data” (data with time spacing less than 24 hours). They released this database last week.

We now have the option of supplementing our existing dataset with this high-quality fine data, which includes two extra years of data (2022 and 2023) collected since we first started work on this paper.

In some cases, the new data replaces old data; in other cases, it adds years we did not have before. If we supplement with the new database up to dec. 2023, we gain an additional 18 stations that have at least 20 years of data (232 -> 250).

If we use this new dataset:

The fine database covers only data from about 1970 to 2024. So the actual findata.

I trust this new dataset far more than the original…

In 2022, when we pulled the data for this paper, NVE had no “perfect” archive for either fine data or daily data. Some archives were ice-reduced and others were not. Some archives were quality controlled and gap-filled, while others were not. Some years and stations were in some databases and not others and vice versa.

To get around this Kolbjørn and I spent a lot of time cross-checking archives and doing our best to manually remove problematic data. But manual removal <<< 2 years of checking by the hydrometerological section

Assets

**USGS reports**

This is the significant body of literature on our method.

Something I didn’t catch before is that the USGS hypothesizes that developing a single parametric regression model, and then re-estimating the model coefficients for each duration, helps keep consistency between durations.

So, perhaps we should also address duration consistency in our results section.

These reports test durations between 1 and 30 days.

To better match existing literature:

* Added durations up to 30 days
* Investigated duration consistency in the median flood predictions

A common challenge in engineering design for retention-based applications is the need for duration-specific design values at ungauged locations. Since these applications focus on total storage capacity, they often require frequency estimates of average streamflow over pre-determined durations. Moving-window analyses (alternatively termed sustained flood flow, N-day flood or flood-duration-frequency analyses) enable frequency analysis of the period with the highest average streamflow over $n$ consecutive hours; see, for example, \citet{lamontagne2012development,kennedy2014methods,lind2020development,barna2023flexible,ouarda2006data,javelle2002development,balocki1994relationships}. Frequency analysis is typically performed separately for each pre-determined duration. Some approaches (e.g., \citet{barna2023flexible,ouarda2006data,javelle2002development}) attempt to scale across durations by fitting an `average' distribution and estimating a scaling parameter. However, this scaling relies on strict assumptions that are sometimes unsupported by empirical analysis, especially when applied over broad regions \citep{barna2023flexible,kennedy2014methods}. Separate frequency analyses for each duration are the more common approach in regional applications and are the focus of this paper.

To extend these estimates to ungauged locations, regression models are typically used. Moving-window analyses often model a wide range of durations (e.g., \citet{lamontagne2012development} models durations from 1 to 30 days) by developing a single best parametric regression model and re-estimating the coefficients for each duration. Relying on a single parametric form is practical and is theorized to help maintain consistency across durations \citep{lind2020development}.

However, this approach assumes that the same parametric form applies to all durations. This may not fully account for the fact that different processes drive high average streamflow over short vs long durations. For example, high average streamflow over short durations is often driven by intense rainfall events, while longer durations are influenced by processes like sustained snowmelt. These processes, and their interactions with catchment descriptors used in regression modeling, may be so fundamentally different that simply re-estimating coefficients may not capture the variations in relationships across durations.

Despite this, the assumption of a single parametric form is rarely challenged due to the complexities of developing and testing regional regression models on many durations. Instead, practitioners sometimes limit event types in moving-window analyses \citep{lamontagne2012development} and/or allow for differing amounts of cross-correlation between stations when re-estimating regression coefficients for each duration \citep{lamontagne2012development,kennedy2014methods,lind2020development}. The latter typically requires a two-step estimation procedure as high cross-correlations can occur between gauging stations in moving-window analyses, particularly at longer durations \citep{kennedy2014methods}.

This paper investigates whether use of a single parametric form limits the predictive performance of a regional regression model applied to multiple durations. We examine whether there are statistically significant differences in consistency, predictive accuracy, reliability, and fitted relationships across different durations and different types of regression models.

Of course, establishing a need for different models at different durations is of limited usefulness if it is not practical to build the models. Developing a regional regression model is a substantial effort. Hydrologic datasets of catchment and climate characteristics often contain a large number of potential—and potentially collinear—predictors, making predictor selection a challenge. Improving model fit often involves selecting appropriate polynomial terms and transformations of the chosen predictors, which can be a labor-intensive process. And models used to produce design values must be tested on a wide variety of performance metrics that assess predictive accuracy, reliability, and uncertainty at ungauged locations.

Therefore, implementation potential played a large role when considering what types of regression models to assess. We considered several options, all of which have established use in regional hydrologic studies.

Fully parametric regression models, like linear, log-linear, or nonlinear models, are the current state of practice in moving window analyses and are well-established in regional flood frequency analysis in general \citep{Robson1999,pandey1999comparative}. These models are straightforward to interpret and assess but require significant work to develop an appropriate parametric form. Additionally, they rely on careful selection of uncorrelated predictors, a major part of the analysis. In contrast, fully data-driven (non-parametric) models, such as the machine learning models of, for example, \citet{aziz2014application, laimighofer2022parsimonious, jarajapu2022design, haddad2020regionalsvm} or \citet{esmaeili2023regional}, are highly adaptable, can handle large, collinear predictor sets, and do not need a defined parametric form. They are powerful methods when predictive accuracy is the only concern. But these models generally lack the distributional assumptions needed for uncertainty and reliability assessments. A middle ground can be found in semi-parametric regression models, such as generalized additive models (GAMs) (see, e.g., \citet{chebana2014regional,msilini2022flood,rahman2018development}), which model predictor-response relationships non-parametrically while still specifying the response distribution, enabling full statistical assessments of accuracy, reliability, and uncertainty.

We develop a GAM for regional median flood prediction at multiple durations. Some moving window analyses use the regression model to predict a distribution parameter for frequency analysis \citep{lamontagne2012development,lind2020development}, while others predict flood quantiles directly \citep{kennedy2014methods}. In this investigative study, we focus on predicting the median annual maximum flood (index flood, as in \citet{Dalrymple1960}) which can serve as both a quantile and, through the reparametrization of \citet{castro2022practical}, the location parameter of the generalized extreme value distribution.

Just as with a standard parametric model in moving window analyses, we specify the GAM once and re-estimate at each duration. However, unlike the parametric model, the GAM’s explanatory component--the predictor-response relationship--adapts as we re-estimate, offering a data-driven comparison and visualization of fitted relationships and estimation uncertainty across durations. For the predictive component of our model assessment, we compare the models' predictions to observed data. The models' predictive power at ungauged locations is assessed through a cross validation study, where reliability and accuracy are evaluated through comparison between predictions and holdout data. \textcolor{red}{can add this if we add consistency:} Finally, predictions at ungauged locations are assessed for statistically significant deviations from duration consistency.

In our context (design values for retention-based applications) reliability and uncertainty analyses are highly relevant. As noted earlier, this limits our reliance on machine learning. However, machine learning models are powerful and popular methods. For the sake of completeness, we include machine learning in our analysis to the extent possible. In addition to the GAM, we develop also a gradient-boosted tree ensemble (XGBoost) for median flood prediction and use it to provide a benchmark value for predictive accuracy. XGBoost has established use in hydrology \citep{zounemat2021ensemble} and is applied in, for example, \citet{laimighofer2022low} and \citet{ni2020streamflow}. Additionally, studies have shown tree-based models like XGBoost to be effective in identifying small, non-redundant predictor sets from larger, sometimes highly correlated, predictor sets (see, e.g., \citet{alsahaf2022framework}, \citet{galelli2013tree}, and \citet{prasad2017input}). We implement the approach of \citet{galelli2013tree} and assess its effectiveness for our specific study.

Our study area is Norway, where we compare the GAM to an existing index flood model for Norwegian catchments. The models are compared on a set of durations ranging from 1 hour to 30 days. Shorter durations are relevant to Norway’s many small hydropower plants, while longer durations apply to larger reservoirs. The focus on annual maxima aligns with Norwegian flood guidelines. The following research questions will be addressed:

\begin{enumerate}

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

\item 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?).

\end{enumerate}

The remainder of the paper is organized as follows: section \ref{sec:data} introduces the flood data and catchment descriptors. Section \ref{sec:study\_design} presents an outline of the study design. Section \ref{sec:methods} 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 \ref{sec:results} presents the predictive performance and model reliability results as well as the functional relationships identified by the GAM. The paper finishes with a discussion (section \ref{sec:discussion}) and conclusions (section \ref{sec:conclusions}).