SUPPLEMENTARY MATERIAL TO

"NEURAL NETWORKS FOR EXTREME QUANTILE REGRESSION WITH AN APPLICATION TO FORECASTING OF FLOOD RISK"

BY OLIVIER C. PASCHE[1,a] AND SEBASTIAN ENGELKE[1,b]

[1]*Research Center for Statistics, University of Geneva, Switzerland,* [a]*olivier.pasche@unige.ch;* [b]*sebastian.engelke@unige.ch*

S.1. Additional LSTM illustration. Figure S.1 shows a schematic representation of a multilayer LSTM network.

S.2. Details on Algorithms 1 and 2. Algorithms 1 and 2 contain some abbreviated function calls. We give some details here:

RANDOMVALIDATIONSPLIT($\mathcal{I}$): For independent data, splits the index set $\mathcal{I}$ randomly into training set $\mathcal{T}$ and validation set $\mathcal{V}$ with prespecified proportions.
SEQUENTIALVALIDATIONSPLIT($\mathcal{I}$): For sequential data, splits the index set $\mathcal{I}$ sequentially into training set $\mathcal{T}$ and validation set $\mathcal{V}$ with prespecified proportions, such that all observations in $\mathcal{T}$ are before $\mathcal{V}$ in time.
INITIALIZENETWORKWEIGHTS(Θ) / INITIALIZERECURRENTNETWEIGHTS(Θ): Initializes the weights of the GPD (recurrent) neural network randomly. The number of weights is determined by Θ.
GETMINIBATCHES($\mathcal{T}$): Splits the training set $\mathcal{T}$ into mini-batches for stochastic gradient descent.
BACKPROPUPDATE(ℓ, $\hat{\mathcal{W}}$, $\mathbf{x}_{\mathcal{B}}$, $\hat{Q}_{\mathbf{x}_{\mathcal{B}}}(\tau_0)$, Θ): Updates the parameter vector by a gradient step computed by backpropagation; may involve regularization methods such as L_2-penalty or dropout, specified in the hyperparameters Θ.
LOSSNOTIMPROVING($\hat{\mathcal{W}}$, $\mathbf{x}_{\mathcal{V}}$, $\hat{Q}_{\mathbf{x}_{\mathcal{V}}}(\tau_0)$, $z_{\mathcal{V}}$): If validation loss is tracked, then also a stopping criterion (e.g., for early stopping) is specified, and this function returns `TRUE` if this criterion is attained, indicating that the validation loss is not improving any more.

S.3. Simulation study for independent observations. Three data-generating models with independent observations are considered. The training sample $\mathcal{D} = \{(\mathbf{x}_i, y_i)\}_{i=1}^n$ is drawn from

$$\begin{cases} \mathbf{X} \sim \mathcal{U}\left([-1,1]^p\right), \\ Y \mid \mathbf{X} = \mathbf{x} \sim \sigma(\mathbf{x}) \cdot t_{\alpha(\mathbf{x})}, \end{cases} \tag{S.1}$$

with $p = 10$, $\alpha(\mathbf{x}) = 1/\xi(\mathbf{x}) := 7 \cdot \{1 + \exp(4x_1 + 1.2)\}^{-1} + 3$, and three different models for $\sigma(\mathbf{x})$:

Model 1: $\sigma(\mathbf{x}) := 1 + 6\phi(x_1, x_2)$, where ϕ is the bivariate Gaussian density with correlation 0.9,
Model 2: $\sigma(\mathbf{x}) := 4 + 3\cos(7 \left\|(x_1, x_2)^\top\right\|_2 + 3)$,
Model 3: $\sigma(\mathbf{x}) := 4 + 3\cos(6 \left\|\mathbf{x}\right\|_2 + 3.5)$.

To avoid bias in the selection of the data models, $\alpha(\mathbf{x})$ and $\sigma(\mathbf{x})$ in Model 1 are the same as in the simulation study in Velthoen et al. (2023). The choices for Models 2 and 3 are designed to study more complex covariate dependencies. The constants are chosen to have positive scale values and enough variation for the inference task to be interesting.

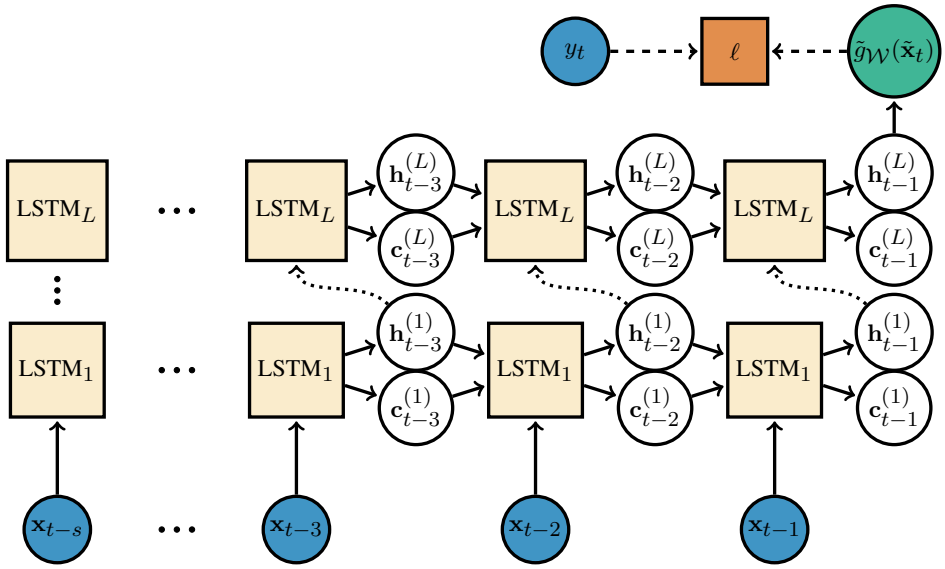

FIG S.1. *Multilayer LSTM network flowchart from input* $\tilde{\mathbf{x}}_t := (\mathbf{x}_{t-s}, \ldots, \mathbf{x}_{t-1})$ *to output* $\tilde{g}_{\mathcal{W}}(\tilde{\mathbf{x}}_t)$*, with loss evaluation. The LSTM cells represent the transformation in* (9)*.*

As the focus of the experiments is on the extremal part of the model and since all extreme value competitors use the same intermediate quantile estimates, we use the true $Q_{\mathbf{x}}(\tau_0)$ as intermediate quantiles. Two main types of network architectures are considered for EQRN. The first type is an MLP with $\tanh$ activation functions, narrow architectures with between one and four hidden layers and optional L_2 weight penalty as training regularization. The second type is self-normalizing networks (Klambauer et al., 2017) using SELU activation functions, deeper architectures with between three and eight hidden layers and optional alpha dropout as regularization. This type of network is designed to maintain unit variance and zero mean across layers in deeper networks trained for regression tasks, in order to avoid vanishing and exploding gradient issues. Results suggest that the additional flexibility of the second type was not necessary for the three tasks at hand, as they never yielded better validation scores than the best $\tanh$ models. Table S.1 summarizes the hyperparameters of the chosen EQRN networks.

To evaluate the accuracy of the best models over the full feature space $\mathcal{X} = [-1, 1]^p$, we use the integrated squared error (ISE) between the prediction $\hat{Q}_{\mathbf{x}}(\tau)$ and the true quantile $Q_{\mathbf{x}}(\tau)$,

$$\int_{\mathcal{X}} \left(\hat{Q}_{\mathbf{x}}(\tau) - Q_{\mathbf{x}}(\tau)\right)^2 d\mathbf{x}. \tag{S.2}$$

We generate test features using a Halton sequence (Halton, 1964) and compute the MSE between the corresponding predicted and true response quantiles, to estimate the p-dimensional integral.

Figure S.2 shows the accuracy of EQRN and the competitor methods for an increasingly large τ, for the three data models. The competitors' performances shown here were also significantly improved by using the intermediate $\hat{Q}_{\mathbf{x}}(\tau_0)$ as an additional covariate. For every model, EQRN outperforms all competitors, with a difference in accuracy increasing with τ. EGAM seems to suffer from the large dimension of the feature space at large τ, both when the actual quantile depends on only two or all of the ten features. The difference in accuracy of EQRN and GBEX compared to the unconditional and semiconditional models is particularly significant for Models 1 and 3, and EQRN generally outperforms GBEX, especially at high probability levels.

TABLE S.1
Hyperparameters of the EQRN networks with the best validation loss for the three independent data models.

	Hidden activation function	Hidden layer dimensions	L_2 penalty
Model 1	tanh	$(128, 128, 128)$	10^{-5}
Model 2	tanh	$(20, 10, 10)$	10^{-5}
Model 3	tanh	$(10, 10, 10)$	0

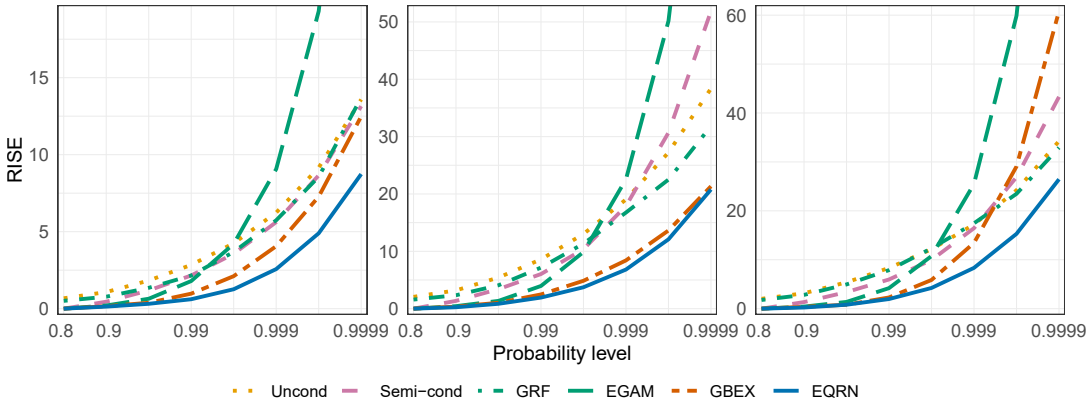

FIG S.2. *Root integrated squared error between predicted and true conditional quantiles at different probability levels τ (log-scale) for the selected EQRN model and the improved competitors, for data Models 1–3 (left to right). The cropped-out RMISE for EGAM at level 0.9999 are around 43, 115 and 150, respectively.*

We also define the quantile R squared of $\hat{Q}_{\mathbf{x}}(\tau)$ over the sample $\mathcal{D}$ as

$$R^2_\tau := 1 - \frac{\sum_{i=1}^n \left(Q_{\mathbf{x}_i}(\tau) - \hat{Q}_{\mathbf{x}_i}(\tau)\right)^2}{\sum_{i=1}^n \left(Q_{\mathbf{x}_i}(\tau) - \overline{Q_{\mathcal{D}}(\tau)}\right)^2}, \quad \text{with} \quad \overline{Q_{\mathcal{D}}(\tau)} := \frac{1}{n}\sum_{i=1}^n Q_{\mathbf{x}_i}(\tau). \tag{S.3}$$

The definition is similar to the classical R squared coefficient of determination in regression, but the true conditional quantile values are used as targets instead of the response observations. The R^2_τ is essentially the reversed MSE normalized by the variance of the true conditional quantile. A value close to unity indicates a very low MSE compared to the quantile variance, and negative values indicate a MSE larger than the quantile variance.

Figure S.3 shows the quantile R squared, the biases and the residual standard deviations of the same respective quantile predictions compared to the truth. The R squared lead to the same conclusions as the RISE. Regarding the bias-variance decomposition of the RISE, it seems that the variance term is dominating the square bias. Although EQRN is here not the least biased model for large τ values, it is its dominating performance in terms of residual variance that leads to it having the lowest RISE values for every data model.

Figure S.4 shows the predicted $\hat{Q}_{\mathbf{x}}(0.9995)$ for EQRN and the competitors, as a function of the two significant covariates (x_1, x_2) for Model 1. At that extreme level, EQRN still seems to capture the true conditional quantile function quite well, although the predictions show some residual noise. GBEX shows an elliptical stepwise approximation behaviour. and seems to underestimate the smaller quantiles and overestimate the largest quantiles. EGAM and GRF fail drastically at recovering the conditional quantile function. The semiconditional estimates are a translation of the intermediate quantiles, which in particular fail to capture the varying shape along X_1.

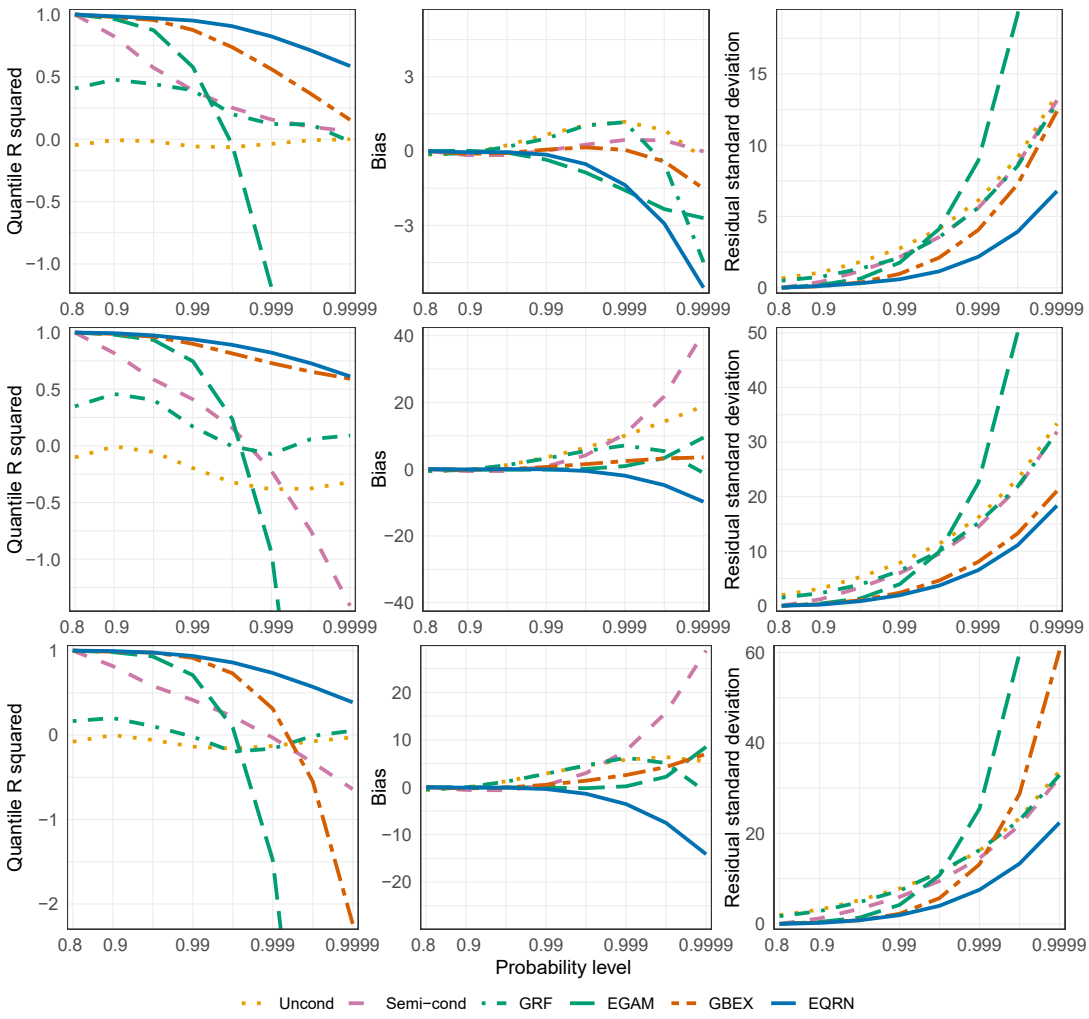

FIG S.3. *Quantile R squared, bias and residual standard deviation of the predicted quantiles compared to the truth at different probability levels (log-scale) for the selected EQRN model and improved competitors, for data Models 1–3 (top to bottom).*

S.4. Simulation study for sequentially dependent data. The main results of the simulation study on sequentially dependent data are presented in the main paper. This section discusses additional results. Figure S.5 shows part of the sequential data simulated from the generating process described in the main paper.

Figure S.6 shows the quantile R squared (S.3), the biases and the residual standard deviations of the quantile predictions compared to the truth, for the two selected EQRN models and competitors. The R squared evolution again shows EQRN is the model that best captures the covariate sequential dependence in the tail, as it outperforms all competitors with a difference in accuracy increasing with τ. In terms of bias, the penalized EQRN here scores similar values as EXQAR, and both EQRN versions outperform all other methods. The unpenalized EQRN has the lowest residual variance, closely followed by both GBEX and the penalized EQRN, although GBEX has a bad accuracy overall, due to its large bias.

Figure S.7 shows the impact of the intermediate level τ_0 on the accuracy of the best EQRN model. The value $\tau_0 = 0.8$ used in the rest of the analysis leads to the best RMSE. This relatively low value shows that τ_0 can be chosen much lower than for the classical

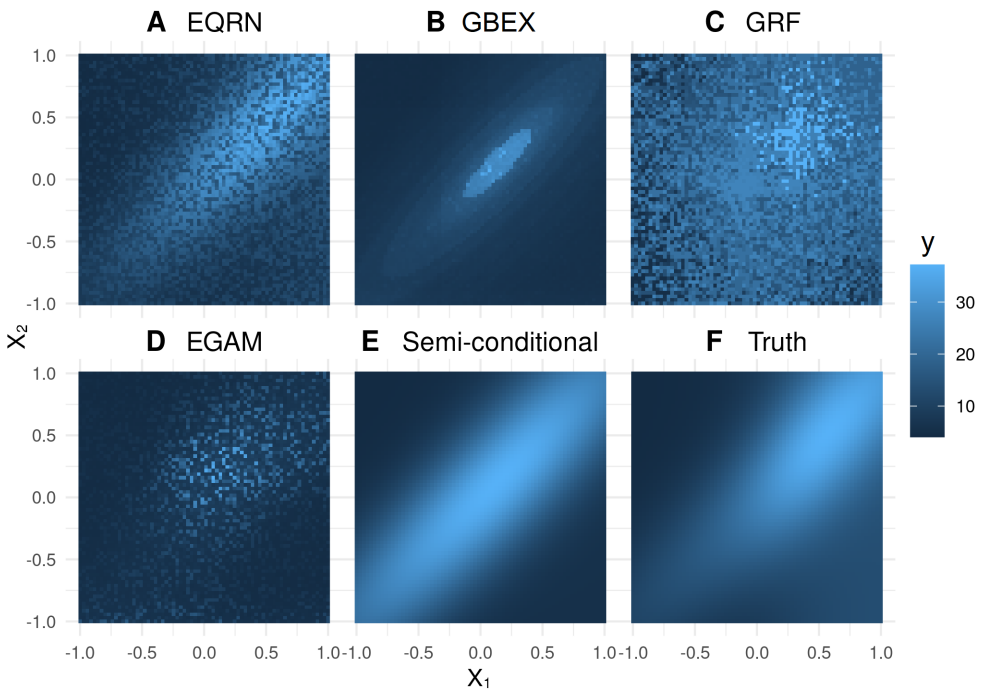

FIG S.4. *Conditional quantile predictions of EQRN and the improved competitor models at probability level* $\tau = 0.9995$*, shown as a function of* X_1 *and* X_2*, for Model 1.*

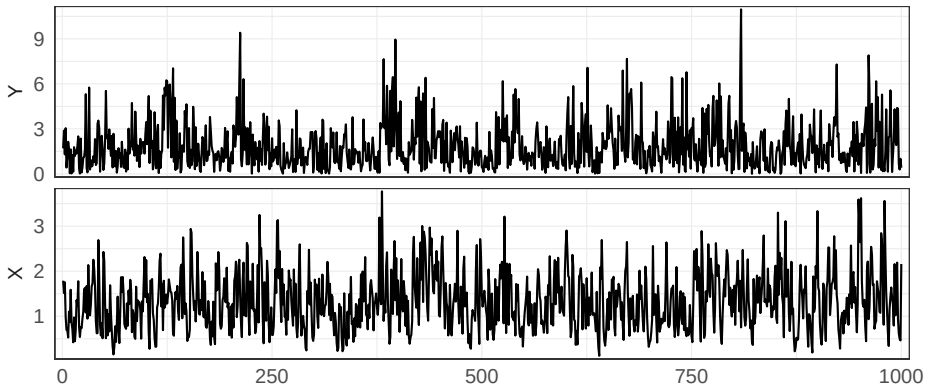

FIG S.5. *First* 1,000 *observations of the sequential data simulated from* (13).

unconditional GPD model. An intuitive explanation for this fact is the following. In a situation without covariates, the choice of τ_0 is a trade-off between approximation bias (which favours larger thresholds) and variance (which favours lower thresholds); see Figure 3 in the main document. For covariate-dependent data, the distribution of the exceedances varies and more data is needed to accurately capture this function of the covariates. The variance becomes more important than the approximation bias, therefore, lower thresholds are preferable. Moreover, the flexibility of the GPD regression neural network model seems to be able to absorb some of the approximation bias, also allowing for a low value of τ_0.

The final accuracy seems in fact to not be too sensitive to the choice for τ_0 compared to the network's grid-searched hyperparameters, discussed in the main analysis, as the differences in RMSE for τ_0 values close to the optimum are relatively small. As mentioned in Section 3 of the main paper, τ_0 cannot be treated as a classical tuning parameter, as different values for τ_0

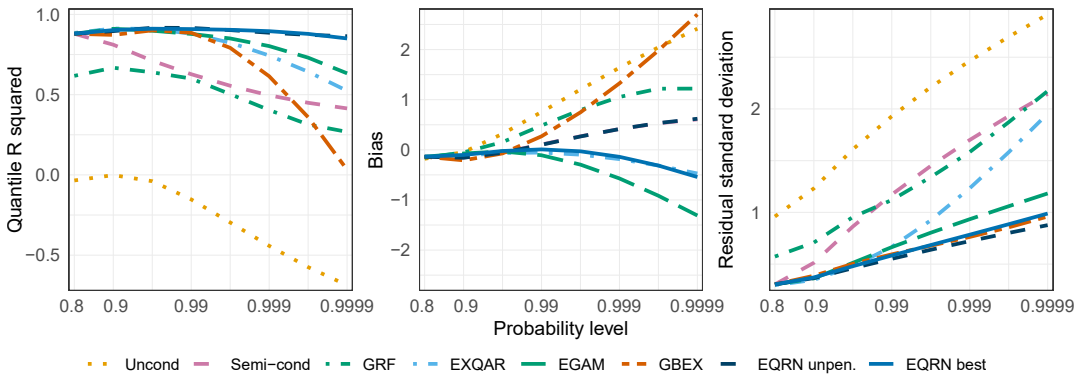

FIG S.6. *Quantile R squared, bias and residual standard deviation of the predicted quantiles compared to the truth at different probability levels (log-scale) for the selected EQRN model and the improved competitors, for the sequential data model.*

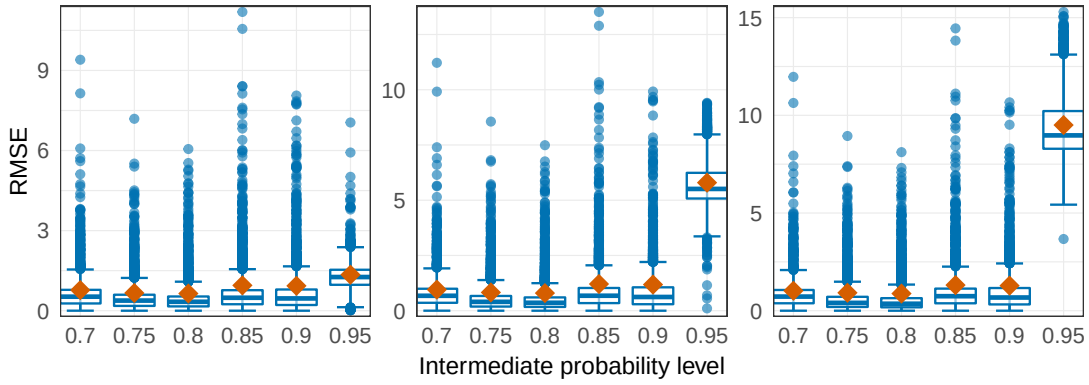

FIG S.7. *Boxplot of the absolute residuals of the quantile predictions from the selected EQRN model (blue) and their RMSE (red diamond) at probability levels* 0.995 *(left),* 0.999 *(middle) and* 0.9995 *(right) for different choices of intermediate probability level* τ_0*, for the sequential data model.*

generally yield different subsets of exceedances. Thus, the likelihood (11), which is used as a goodness of fit metric for hyperparameter tuning, would not be comparable between models. The RMSE cannot be computed in practice either, since $Q_{\mathbf{x}}(\tau)$ is generally unknown.

S.5. Application: competitor results. The main results from our application to forecasting flood risk in Switzerland using our proposed EQRN methodology are presented in the main paper. This section discusses and compares additional results using the competitor methods, adapted to provide the same type of forecast as the EQRN approach, with a focus on the 2005 flood event (see Figure 2 in the main paper).

We first observe that the predictions have roughly the same behaviour, which shows that all methods capture at least some of the temporal structure based on the past covariates. The semiconditional method (Figure S.8) is clearly not flexible enough, since the predictions only show a weak sensitivity to the changes in covariates. It also fails to trigger any early warning during the main event, due to low probability ratio forecasts never exceeding the selected threshold value of 100. The reason is that, while the intermediate quantile is covariate-dependent, the GPD parameters are constant over time. We conclude that a covariate-dependent model for the tail is required in this application. Figure S.9 shows predictions from the EXQAR

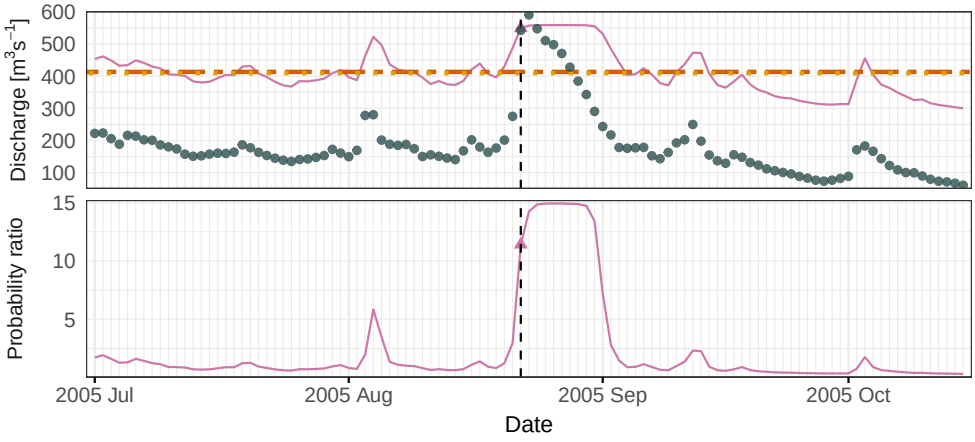

FIG S.8. *Top: Daily average discharge (points) at the Bern–Schönau station (62) and one-day-ahead semiconditional forecasts of conditional 100-year quantiles (solid line) during the 2005 flood. Horizontal lines show unconditional Q^{100} based on GEV (dashed) and GPD (dotted). Bottom: One-day-ahead forecast of the conditional probability of exceeding the GEV estimated Q^{100} as a ratio to the unconditional probability, using the semiconditional parameter forecast. The vertical line indicates August 22, the day of the first exceedance.*

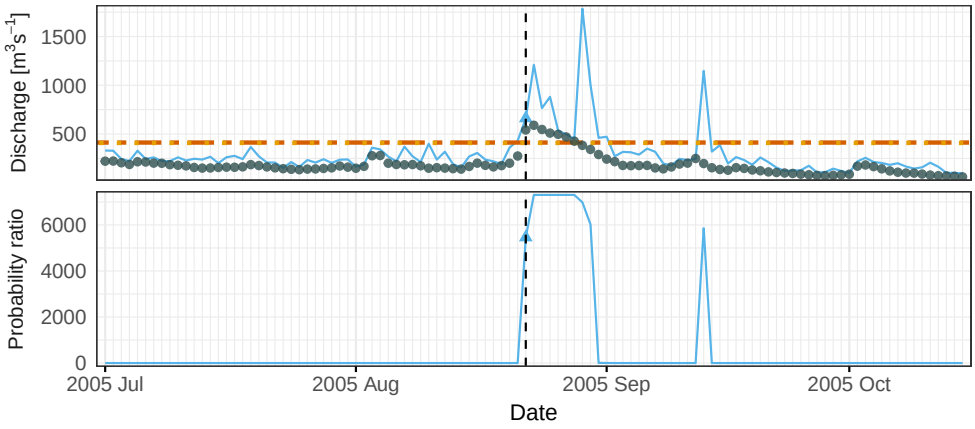

FIG S.9. *Top: Daily average discharge (points) at the Bern–Schönau station (62) and one-day-ahead EXQAR forecasts of conditional 100-year quantiles (solid line) during the 2005 flood. Horizontal lines show unconditional Q^{100} based on GEV (dashed) and GPD (dotted). Bottom: One-day-ahead forecast of the conditional probability of exceeding the GEV estimated Q^{100} as a ratio to the unconditional probability, using the EXQAR forecast. The vertical line indicates August 22, the day of the first exceedance.*

model (Li and Wang, 2019). This model is more sensitive to changes in the covariates, but the regression function looks fairly erratic, with sudden spikes at some time points. Those spikes are here due to unusually small shape estimates in combination with a large-scale estimate. This might be caused by an instability in the estimation of the local moments used in the model. EGAM (Figure S.10) fails to trigger an early warning for the first day of the flooding event as its quantile and probability ratio forecasts are very low, but then seems to severely overestimate the river flow during the rest of the event. The best competing model seems to be the GBEX (Figure S.11), as it yields a smooth prediction curve with a pronounced spike at the main event. Comparing this with our EQRN method (Figure 2 in the main paper), it reacts slightly later and its risk forecast (probability ratio) on the day before the first exceedance is significantly lower than for EQRN.

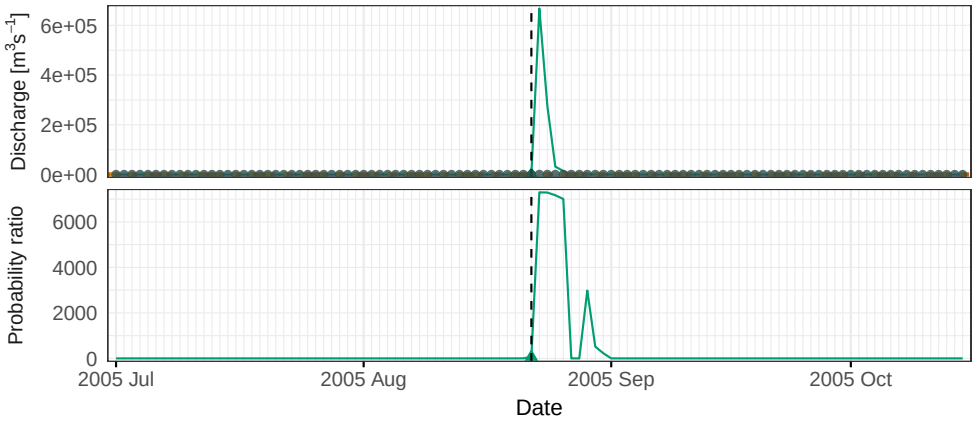

FIG S.10. *Top: Daily average discharge (points) at the Bern–Schönau station (62) and one-day-ahead EGAM forecasts of conditional 100-year quantiles (solid line) during the 2005 flood. Horizontal lines show unconditional* Q^{100} *based on GEV (dashed) and GPD (dotted). Bottom: One-day-ahead forecast of the conditional probability of exceeding the GEV estimated* Q^{100} *as a ratio to the unconditional probability, using the EGAM parameter forecast. The vertical line indicates August 22, the day of the first exceedance.*

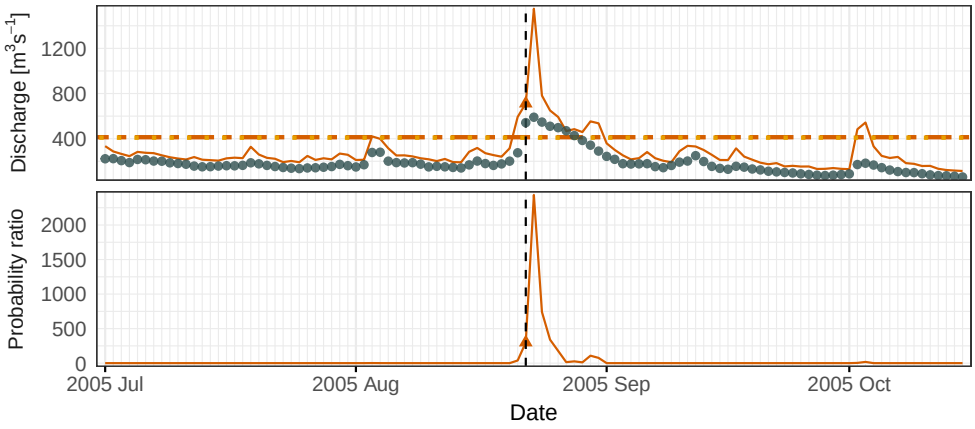

FIG S.11. *Top: Daily average discharge (points) at the Bern–Schönau station (62) and one-day-ahead GBEX forecasts of conditional 100-year quantiles (solid line) during the 2005 flood. Horizontal lines show unconditional* Q^{100} *based on GEV (dashed) and GPD (dotted). Bottom: One-day-ahead forecast of the conditional probability of exceeding the GEV estimated* Q^{100} *as a ratio to the unconditional probability, using the GBEX parameter forecast. The vertical line indicates August 22, the day of the first exceedance.*

This qualitative way of checking the model is important since in real-world applications, the true extreme quantiles are unknown. Therefore, only some quantitative model checks can be performed, like the right-hand panel of Figure 7 in the main paper showing calibration in the number of quantile-exceeding test observations. Figure S.12 compares this number of quantile-exceedances for the competitor predictions. Although they can give evidence against the suitability of a model, it highlights that such metrics only assess calibration but not goodness-of-fit nor accuracy, as unconditional methods obtain similar values to flexible accurate methods. This underlines the importance of the simulation studies, which allow us to evaluate in several settings which methods are more accurate. In particular, in situations with temporal dependence, our EQRN method outperforms the competitors (e.g., Figure 5 in the main paper).

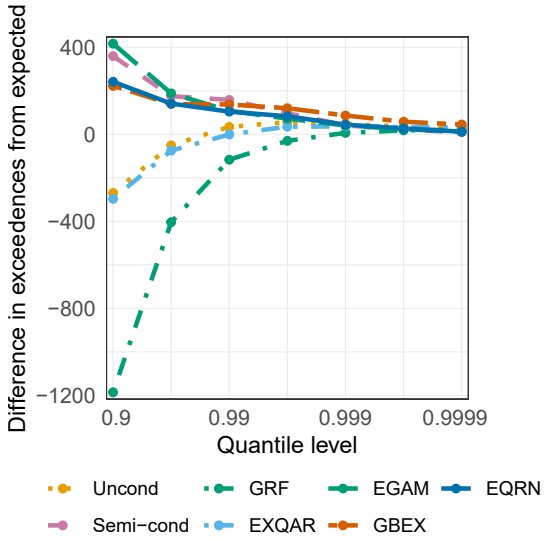

FIG S.12. *Difference between the number of observations exceeding the EQRN and competitor quantile predictions on the test set and the expected number of exceedances, for different probability levels (log-scale).*

REFERENCES

HALTON, J. H. (1964). Algorithm 247: Radical-Inverse Quasi-Random Point Sequence. *Commun. ACM* **7** 701–702.

KLAMBAUER, G., UNTERTHINER, T., MAYR, A. and HOCHREITER, S. (2017). Self-Normalizing Neural Networks. In *Proc. 31st Int. Conf. Neural Inf. Process. Syst.* 972–981.

LI, D. and WANG, H. J. (2019). Extreme Quantile Estimation for Autoregressive Models. *J. Bus. Econ. Stat.* **37** 661–670.

VELTHOEN, J., DOMBRY, C., CAI, J.-J. and ENGELKE, S. (2023). Gradient boosting for extreme quantile regression. *Extremes* **26** 639–667.