Royal Netherlands
Meteorological Institute
Ministry of Infrastructure
and Water Management

Estimation of wind speeds with very high return periods from large datasets generated by weather prediction models : statistical aspects

C.F. de Valk, H.W. van den Brink

De Bilt, 2020 | Scientific report; WR 2020-01

Estimation of wind speeds with very high return periods from large datasets generated by weather prediction models: statistical aspects

C.F. de Valk, H.W. van den Brink

De Bilt, 2020 | Scientific report; WR 2020-01

Samenvatting

Voor de toetsing van de veiligheid van waterkeringen in Nederland zijn waarden van windsnelheid en zeewaterstand voor terugkeertijden tot enkele miljoenen jaren nodig. Dit is een uitdaging, gegeven dat reeksen van betrouwbare wind metingen niet verder teruggaan dan ongeveer 70 jaar.

Momenteel worden verschillende ideeën om dit probleem op te lossen verkend. Eén idee is om het data volume te vergroten door gebruik te maken van omvangrijke datasets van simulaties door numerieke weermodellen; zie van den Brink (2018). Echter, zelfs met gebruik van grote datasets zoals het archief van ECMWF ensemble seizoensvoorspellingen blijft er een groot verschil in terugkeerperiode dat moet worden overbrugd. Een ander idee is het gebruik van modellen van de staarten van verdelingsfuncties die specifiek ontworpen zijn voor extrapolatie over een breed terugkeerperiode-bereik: de Gegeneraliseerde Weibull (GW) staart en de breder toepasbare log-Gegeneraliseerde Weibull (log-GW) staart, met de Weibull staart als bijzonder geval van beide.

Voor deze modellen alsmede voor twee klassieke staart-modellen, de Gegeneraliseerde Pareto (GP) staart en de exponentiële staart, vergelijken we schattingen van de staart van de kansverdeling van windsnelheid bepaald uit 72-jaar reeksen getrokken uit de ECMWF System-4 ensemble seizoensvoorspellingen. De gebruikte data zijn voor een lokatie in de centrale Noordzee. Eén check beschouwt de statistiek van de hoogste waarde in een 72-jaar reeks zoals voorgesteld in van den Brink and Können (2008). Daarnaast bekijken we voor ieder van de modellen de nauwkeurigheid van schattingen van de 10^7-jaar windsnelheid op basis van 72-jaar reeksen, met als referentie schattingen op basis van de volledige dataset. Gebaseerd op deze resultaten is vervolgens de bias geschat in the extrapolaties op basis van de volledige dataset en op basis van een 72-jaar reeks. Dezelfde analyse is uitgevoerd op de meer recente SEAS5 seizoensvoorspellingen, alsmede op de jaarmaxima van windsnelheid uit een groot aantal runs over 1981-2009 met het klimaatmodel Speedy.

De verschillende datasets geven sterk verschillende resultaten. Met name wijkt de distributiefunctie van de SEAS5 windsnelheid aanzienlijk af van die van de System-4 windsnelheid: de SEAS5 windsnelheid is gemiddeld lager, maar met een zwaardere staart, resulterend in veel hogere schattingen van terugkeerwaarden van windsnelheid. Bovendien is de staart van SEAS5 windsnelheid minder regulier dan de staart van System-4 data, waardoor schattingsfouten aanzienlijk groter zijn. Dit vraagt om nader onderzoek van de oorzaken van dit verschil: potentiële bias geassocieerd met de modelformulering lijkt de onzekerheid te domineren.

Vergelijking van schattingen op basis van verschillende modellen van de staart van de windsnelheidsverdeling toont dat de klassieke GP staart, de exponentiële staart en de 1-parameter Weibull staart een aanzienlijke bias kunnen hebben, afhankelijk van de beschouwde dataset. De GW staart voldoet overall het beste; de 1-parameter Weibull staart kan betere schattingen opleveren indien deze stabiel zijn als functie van de drempel.

Alles bij elkaar genomen wijzen de resultaten er op dat schatting van de windsnelheid voor een terugkeertijd tot 10^7 jaar op basis van grote model-gegenereerde datasets zoals System-4 en SEAS5 goed te doen is (met RMS fout kleiner dan 2

m/s), afgezien van de gevonden systematische verschillen tussen deze datasets.

De resultaten van de huidige studie kunnen nu al van nut zijn om de terugkeerwaarden van windsnelheid die momenteel gebruikt worden in de toetsing van primaire waterkeringen te evalueren en zo mogelijk te verbeteren. Aanbevolen wordt om op de GW staart gebaseerde schattingen uit meetgegevens op verschillende locaties te vergelijken met schattingen gebaseerd op de GP en exponentiële staarten, waarbij met name de onzekerheid van de schattingen wordt onderzocht. De analyse van onzekerheid zou ook het effect van variabiliteit over tijdschalen van meerdere jaren moeten omvatten, waar tot nu toe vrijwel geen aandacht aan is besteed.

Dit onderzoek is uitgevoerd voor Rijkswaterstaat en het KNMI MSO project "Towards future climate proof statistical methods for KNMI products on extremes". Wij danken Marcel Bottema en Pieter van Gelder voor de reviews van dit document.

Summary

To assess the reliability of flood protection in the Netherlands, return values of wind speed and coastal water level for return periods up to several million years are needed. This is a major challenge, given that records of reliable wind measurements do not go back further than about 70 years.

Several ideas are currently explored to tackle this problem. One idea is to increase data volume by utilizing large datasets of simulations by numerical weather prediction models; see van den Brink (2018). However, even large datasets such as the archived ECMWF seasonal ensemble forecasts leave a considerable gap in return period to be overcome. Another idea is to use models of the tails of distribution functions which are specifically designed for extrapolation over a wide range of return periods: the Generalized Weibull (GW) tail and the more widely applicable log-Generalized Weibull (log-GW) tail, with the 1-parameter Weibull tail as a special case of both.

For these models and for two classical tail models, the Generalized Pareto (GP) tail and the exponential tail, we compared estimates of the tail of the wind speed distribution derived from subsets of the ECMWF System-4 seasonal ensemble forecast wind speeds for a location in the central North Sea. One check concerns the statistics of maxima over subsamples as proposed in van den Brink and Können (2008)In addition, we checked the accuracy of each of these models in estimating the 10^7-year wind speed from 72-year subsets of the data, using estimates from the full dataset as reference. Based on the results of this check, estimates of worst case bias in extrapolations were made for extrapolations from the full data set as well as from a 72 year subset. These analyses were repeated on the more recent SEAS5 seasonal forecast data as well as on annual maxima of wind speed from a large number of runs over 1981-2009 of the climate model Speedy.

The three datasets give starkly different results. In particular the wind speed distribution of SEAS5 differs considerably from distribution of System-4 wind speed: wind speeds are lower overall, but the tail is heavier, resulting in much higher estimates of return values. Moreover, the SEAS5 tail is less regular than the System-4 tail, resulting in larger estimation errors. This calls for a further investigation of the cause of this difference: the uncertainty appears to be dominated by potential

bias related to the model formulation.

Comparison of estimates based on different models of the tail reveals that the classical GP tail, the exponential tail and the 1-parameter Weibull can be severely biased, depending on the dataset. Overall, the GW tail performs best; the 1-parameter Weibull tail can give better estimates if these are stable as a function of threshold.

Taken together, the results indicate that accurate estimation of the wind speed for a return period up to 10^7 year from large model datasets such as System-4 and SEAS5 is feasible (with RMS error below 2 m/s), provided that systematic differences between these datasets can be resolved.

The outcomes of the present study can already be helpful to assess and possibly improve the return values of wind speed currently in use to assess flood safety in the Netherlands. It is recommended that GW-tail based estimates from measurements at different sites are compared to estimates based on the GP and exponential tails, addressing in particular the uncertainty of estimates. The analysis of uncertainty should also address the effects of interannual variability, which has been largely ignored until now.

This research was carried out for Rijkswaterstaat and the KNMI MSO project "Towards future climate proof statistical methods for KNMI products on extremes". We thank Marcel Bottema and Pieter van Gelder for their reviews of this document.

Contents

List of Figures

List of Tables

Part I. Report

1 Introduction

To approximate the annual probabilities of failure of selected stretches of dike ("dijktrajecten") in the Netherlands, extreme value statistics of load variables such as wind speed and water level are needed for return periods up to several million years.

Estimates of wind speed for such high return periods derived from available measurement data are likely to be very uncertain, as the period covered by reliable wind measurements in the Netherlands is less than 70 years. As an alternative, van den Brink (2018) proposed to use datasets generated by numerical weather prediction models, which may be much larger. An important potential resource is the archive of ECMWF seasonal ensemble forecast data. The latest generation of seasonal forecast data is SEAS5; we also consider System-4 data as these have already been studied quite extensively, see van den Brink (2018). Including the reforecasts, each dataset contains at each grid point effectively between 5000 and 6000 years of weather data representative of the climate over approximately the past 30 years. In addition, an even larger dataset representing over 300,000 years was recently created using the low-dimensional climate model Speedy (Molteni, 2003; Kucharski et al, 2006).

The large sizes of these datasets will likely make it easier to estimate return values for very large return periods. However, the results will be affected by model bias, which will need to be assessed and if possible, corrected. Furthermore, to do this for a return period of more than a million years from a dataset of order 5000 years, extrapolation over several orders of magnitude of return period is still required. This report investigates how we can do the extrapolation, and what level of accuracy can be achieved.

For the extreme value statistics of load variables like wind which are currently prescribed for assessment the reliability of primary flood defense in WBI-2017, different models of the tail are used for different load variables; see Chbab (2017). For wind en coastal water level, these models fit within the framework of classical extreme value theory: the Generalized Extreme Value distribution for maxima over random samples (in practice, often annual maxima), the Generalized Pareto (GP) distribution for peaks-over-thresholds, and sub-classes of these families of limiting distribution functions. An effort was made to avoid overestimation of the uncertainty. For coastal water level, use was made of information from a study of coastal water levels simulated from the System-4 seasonal forecast data for thus purpose. For wind speed, the exponential tail was fitted to measured wind data, which gives likely conservative, but rather precise estimates, and reasonable assumptions were made to bound the uncertainty resulting from lack of knowledge of the true shape parameter of the GP tail.

By now, an alternative to classical extreme value theory is available which is specifically formulated for the purpose of extrapolation over a wide range of return periods without having to make restrictive assumptions. For univariate probability distributions, this is the log-Generalized Weibull (log-GW) tail model, with the Generalized Weibull (GW) tail as optional model for light tails; see de Valk (2016a).

Simulations in de Valk & Cai (2018) indicate that bias in estimates of very high quantiles can be reduced considerably. In an application to coastal water level data from Hoek van Holland, the uncertainty in estimates of the $10,000$ year water level derived from a fitted log-GW model was much lower than of estimates from a fitted GP model.

The present study compares uncertainties in estimates of the 10^7-year wind speed based on different models of the tail. The large sets of wind data from the System-4 and SEAS5 seasonal forecast archives and the Speedy runs serve as a testbed for comparing the different models. At the same time, we try to assess what level of bias and standard error of these estimates is achievable with datasets of the size of a typical time series of wind measurements, and with datasets as large as the seasonal forecast data.

2 Statistical background, models and estimation

Annual maxima, return periods, return values and quantiles The of a load variable such as wind speed is linked to its of exceedance, usually defined as the reciprocal of the annual probability of exceedance. Suppose that a regularly sampled wind speed record contains n values $X_1, ..., X_n$ in a year, then the annual maximum is

$$X_{\max,n} := \max(X_1, ..., X_n)$$

and the return value x_R of wind speed for the return period R satisfies

$$\mathbb{P}(X_{\max,n} > x_R) = 1/R. \tag{1}$$

Suppose for a moment that $X_1, X_2, ...$ form a stationary sequence, so all X_i have the same distribution function F. Then under a rather weak assumption on the serial dependence, it can be shown that for any sequence $z_1, z_2, ...$ such that

$$n(1 - F(z_n)) \to \tau$$

for some $\tau > 0$,

$$\mathbb{P}(X_{\max,n} \leq z_n) \to e^{-\tau\alpha}$$

for some number $\alpha \in (0, 1]$, called the extremal index Leadbetter et al (1983).

For large n (say one or multiple years), this gives the approximation

$$\mathbb{P}(X_{\max,n} > z_n) \approx 1 - e^{-\tau\alpha} \approx 1 - e^{-\alpha n(1-F(z_n))}. \tag{2}$$

.

If furthermore τ is small, then $1 - e^{-\tau\alpha} \approx \tau\alpha$ and therefore,

$$\mathbb{P}(X_{\max,n} > z_n) \approx \tau\alpha \approx \alpha n(1 - F(z_n)).$$

Therefore, if the return period R is large, we have (see (1))

$$1 - F(x_R) \approx (R\alpha n)^{-1}. \tag{3}$$

Defining the upper quantile $Q(p)$ of the instantaneous distribution function F for a probability p as

$$Q(p) := F^{-1}(1-p) \tag{4}$$

(so assuming that F is continuous, $\mathbb{P}(X > Q(p)) = 1 - F(Q(p)) = p$), combining (3) and (4) gives

$$x_R \approx Q(p_R) \tag{5}$$

with

$$p_R := (R\alpha n)^{-1}. \tag{6}$$

It can be shown that the expressions above do not change when the wind speed statistics depend on the season; effectively, F is replaced by its annual average.

The approximation (5) relates the return value for a return period R to a quantile of the mean instantaneous distribution function F at a probability p_R. It allows us derive return values from an estimate of the tail of F based on all wind speed data exceeding some threshold. This requires an estimate of the extremal index α. The estimator of α from Ferro & Segers (2003) was applied to the 6-hourly System-4 seasonal forecast wind speed data. With increasing threshold, the estimates tend to 1, the same value as for an independent sequence. Therefore, $\alpha = 1$ has been used in all subsequent analyses.

Models of the tail of a distribution function In the present study, four models of the upper tail of a distribution function F are considered. They are all "tail limits": approximations which are assumed to become more accurate with increasing wind speed.

Two models are classical: the Generalized Pareto (GP) tail limit, and the exponential tail limit, which is a special instance of the GP limit. The GP tail limit is de Haan & Ferreira (2006)

$$\lim_{p \downarrow 0} \frac{1 - F(Q(p) + xa(p))}{p} = (1 + \gamma x)^{-1/\gamma} \tag{7}$$

for some positive function a (the local scale parameter), and some real number γ which we will refer to as the shape parameter[1]. The term $Q(p)$ in the argument of F can be regarded as a local offset.

In essence, this model is nonparametric: the mathematical form of the right-hand side of (7) is not specified, but can be derived (see de Haan & Ferreira, 2006)[2].

[1] In statistics, is known as the extreme value index.

[2] In Dillingh et al (1993) and in later reports referring to Dillingh et al (1993), a distinction is made between a nonparametric (VVM-c or VVM-0) model, and a parametric GP model for the tail above some high threshold. Modern treatments such as de Haan & Ferreira (2006) make clear that when applied above a high threshold, estimators based on parametric approximation (such as the maximum likelihood estimator (MLE) or the method of moments) may have similar properties as nonparametric methods, under similar assumptions. The reason is that the GP tail limit is in essence a nonparametric model. It should be kept in mind, however, that as the GP tail is only an approximation, the MLE does not have the same properties as when the tail would be described exactly by a GP distribution above the threshold.

It leads directly to the approximation

$$1 - F(z) \approx p \left(1 + \gamma \left(\frac{z - Q(p)}{a(p)}\right)\right)^{-1/\gamma} \tag{8}$$

for large z.

The exponential tail limit is the limiting case of (7) when $\gamma \to 0$: then its right-hand side becomes e^{-x}, and for large z, (8) becomes

$$1 - F(z) \approx p e^{(Q(p)-z)/a(p)}.$$

The other three models are nonclassical. They are designed to approximate a ratio of logarithms of probabilities instead of a ratio of probabilities as in (7) (see de Valk (2016a,b)). This makes it possible to approximate the distribution function over a wide range of probabilities, although the approximation is a crude one. The first one is the Generalized Weibull (GW) tail limit:

$$\lim_{p \downarrow 0} \frac{\log\left(1 - F(Q(p) + xf(p))\right)}{\log p} = (1 + \rho x)^{1/\rho}, \tag{9}$$

for some positive function f (the local scale parameter), with ρ the shape parameter. Note that (9) has a similar form as (7), but probabilities have been replaced by their logarithms. For large z, it leads to the approximation

$$1 - F(z) \approx p^{\left(1+\rho\left(\frac{z-Q(p)}{f(p)}\right)\right)^{1/\rho}}.$$

A closely related model is the log-GW tail limit, which assumes that the logarithm of the random variable concerned satisfies a GW tail limit. This gives

$$\lim_{p \downarrow 0} \frac{\log\left(1 - F(Q(p)e^{xg(p)}\right)}{\log p} = (1 + \theta x)^{1/\theta} \tag{10}$$

for some positive function g (the local scale parameter), with θ the shape parameter. For large z, it leads to the approximation

$$1 - F(z) \approx p^{\left(1+\theta\left(\frac{\log z - \log Q(p)}{g(p)}\right)\right)^{1/\theta}}.$$

For random variables which are positive with probability greater than 0, the log-GW model is more widely applicable than the GW model. The GW model applies to relatively light tails, not much heavier than an exponential tail. The log-GW model applies also to fat tails such as tails satisfying a GP tail limit with $\gamma > 0$ (power-law distributions). The Weibull tail limit is a special case of the log-GW tail limit, obtained by setting $\theta = 0$ in (10). This gives

$$\lim_{p \downarrow 0} \frac{\log\left(1 - F(Q(p)e^{xg(p)}\right)}{\log p} = \exp(x) \tag{11}$$

and for large z, the (1-parameter) Weibull tail approximation

$$1 - F(z) \approx p^{(z/Q(p))^{1/g(p)}} = \exp\left((z/Q(p))^{1/g(p)} \log p\right).$$

If $\rho > 0$ in the GW tail limit (9), then it is equivalent to a Weibull tail limit (11) with scale g satisfying that $\lim_{p\downarrow 0} g(p) = \rho$; setting $f = Q\rho$ in (9) gives (11) with $g(p) = \rho$. The functional form of the 1-parameter Weibull tail as defined here is compatible with (but more restrictive than) the conditional 2-parameter Weibull distribution of exceedances of a threshold which is frequently used by offshore engineers for extreme value analysis.

The functional forms of all the basic tail limits (the right-hand sides of (7), (9), (10)) are not imposed; they are simply the only possible limits (the exponential and 1-parameter Weibull tail limits being special cases). It should be noted that the approximations indicated by $\approx$ above for the GW and log-GW (incl. Weibull) models are of a different nature than the approximations for the GP (incl. exponential) model: the GW and log-GW models provide a crude approximation applicable over a wide range of probabilities, whereas the GP offers a more accurate approximation applicable over a narrow range of probabilities.

For wind speed, there are reasons to expect that all tail limits listed above apply. For many years, it has been common knowledge that empirical distribution functions of midlatitude wind speed are well approximated by a Weibull distribution; see e.g. Justus et al (1976). The Weibull distribution satisfies the Weibull tail limit (11) with constant scale g, and therefore, also the log-GW tail limit (10) with $\theta = 0$, and the GW tail limit (9) with $\rho = g$. Furthermore, the Weibull satisfies the classical GP tail limit (7) with $\gamma = 0$ (the exponential tail limit). There is of course no reason to assume that the Weibull distribution would fit the tail section of the wind speed distribution as closely as its bulk section. However, outside the known regions affected by tropical cyclones and squalls, neither would one expect the tail section to deviate much from the bulk section. In fact, there is ample empirical evidence that a Weibull tail can provide a good approximation of the tail of the wind speed distribution; see e.g. Cook (1982); Harris (2005); van den Brink and Können (2008, 2011).

Approximations of return values Approximations of return values of wind speed (defined by (1)) for a large return period R follow from the formulas above. For the GP model, the approximation is (see (5), (4) and (6)):

$$x_R \approx Q(p_R) \approx Q(p) + a(p)\frac{1}{\gamma}\left(\left(\frac{p}{p_R}\right)^{\gamma} - 1\right). \tag{12}$$

In the special case of the exponential model ($\gamma = 0$), this reduces to

$$x_R \approx Q(p) + a(p)\log\left(\frac{p}{p_R}\right). \tag{13}$$

For the GW model, the approximation of the return value is

$$x_R \approx Q(p) + f(p)\frac{1}{\rho}\left(\left(\frac{\log p_R}{\log p}\right)^{\rho} - 1\right), \tag{14}$$

and for the log-GW model, it is

$$x_R \approx Q(p)e^{g(p)\frac{1}{\theta}\left(\left(\frac{\log p_R}{\log p}\right)^{\theta} - 1\right)}, \tag{15}$$

which in the special case of $\theta = 0$ (1-parameter Weibull model) reduces to

$$x_R \approx Q(p)\Big(\frac{\log p_R}{\log p}\Big)^{g(p)}. \tag{16}$$

Estimators Estimators of return values from a sample of data $X_1, ..., X_n$ employ the approximation formulas (12)-(15), with $p = k/n$ $(1 \leq k \leq n)$ and $Q(p)$ replaced by its estimate $X_{n-k+1:n}$, the $k-$th highest value from $X_1, ..., X_n$.

Estimation of the scale parameter ($a(p)$, $f(p)$ or $g(p)$) and the shape parameter (γ, ρ or θ) in these formulas is less simple. For the GP model; the maximum likelihood estimator (MLE) was chosen see e.g. Section 3.4 of de Haan & Ferreira (2006) for the background. For the log-GW and GW models, a refinement of the method described in de Valk & Cai (2018) was used; see Appendix B. The exponential model was treated as a special case of the GW model, so the same estimator was used with GW shape parameter fixed at 1. The 1-parameter Weibull tail was treated as a special case of the log-GW tail with shape parameter fixed at 0; the scale estimator for this tail is almost identical to the estimator denoted by $\hat{\theta}^{(2)}$ in Gardes and Girard (2016).

Code of the estimators in the R language is contained in the package EVTools, which can be found on https://github.com/ceesfdevalk/EVTools.

One aspect of these estimators which is important for understanding the data analysis in the next chapters is the choice of thresholds. Each estimator for a scale or shape parameter operates on the highest k values in the data sample, i.e., on $X_{n-k+1:n}, ..., X_{n,n}$, for some choice of $k < n$. Choosing a small value of k reduces bias, whereas a large value of k reduces variance. The estimators for the GW and log-GW tails use a larger value of k for the shape parameter than for the scale parameter; this is needed to ensure consistency without imposing restrictive assumptions (see also Appendix B). For the GP tail (estimated by maximum likelihood), this is not needed, but there is also no reason to think that estimation with different thresholds for scale and shape would not work in this case. Therefore, we compare estimates with different thresholds for scale and shape for all tail models, and we compare estimates with the same thresholds for scale and shape for all tail models. In this way, we do not confuse the effects of the choice of tail model with the effects of threshold choice.

3 Analysis of System-4 data

3.1 System-4 data and preprocessing

ECMWF computed System-4 forecast ensembles starting at the first day of every month in 1981-2016, with each forecast run over 7 months. Of every forecast, we retained only the last 6 months in order to minimize the effect of the initialization, ensuring that dependence among the ensemble members is sufficiently reduced. In the reforecast data, ensemble size is 15, except when starting in Feb, May, Aug and Nov; then the ensemble size is 50. In the forecast data (starting from 2011), the ensemble size is always 50, but the additional ensemble members were skipped in order to ensure that all years are equally represented. The time step is 6 hours.

Further information about the System-4 data can be found in Molteni et al (2011); see van den Brink and de Goederen (2017) for a statistical application.

For the present study, System-4 data of wind at 10m above the surface were selected for the location 55N, 3E in the central North Sea; this choice reduces possible effects of land-sea boundaries on the tail of the distribution function of wind speed.

Subsamples of the data were generated by combining ensemble members with the same label and with starting dates at regular 6-month intervals. This resulted in 160 subsamples covering 36 years each. Pairs of these were combined, resulting in 80 subsamples, each covering 72 years: slightly longer than the longest quality-controlled measurement record of wind in the Netherlands.

3.2 Statistics of the highest wind speed in a subsample

The first check of the extrapolation skills of the tail models of Section 2 is taken from van den Brink and Können (2008, 2011). With a chosen model, we fit the tail of each of the 80 subsamples of the seasonal wind speed forecasts (see Section 3.1). For a given subsample, this tail fit allows us to transform the wind speed data monotonically to make their tail distribution standard exponential. If the tail fit matches the true tail closely, then true distribution function of the transformed wind speed should be close to the standard exponential distribution even at the very far end of the tail. Therefore, Z, the highest transformed wind speed from the subsample (the "outlier"), should be approximately Gumbel distributed:

$$\mathbb{P}(Z \leq x) \approx e^{-n\alpha e^{-x}},$$

with α the extremal index and n the number of data points in the subsample. This follows directly from eq. (2). Equivalently, the distribution function of $Y := Z - \log(n\alpha)$ (from now on referred to as the transformed wind speed maximum) should be close to a standard Gumbel distribution:

$$\mathbb{P}(Y \leq x) \approx e^{-e^{-x}}. \tag{17}$$

Since we have 80 subsamples, we can check whether the approximation (17) is true: the values of Y from each of the 80 subsamples, say $Y_1, ..., Y_{80}$, can be sorted to obtain the order statistics $Y_{1:80} \leq ... \leq Y_{80:80}$. According to (17), the plot of $Y_{i:80}$ against $-\log(-\log(i/81))$ (the Gumbel probability scale) for $i = 1, ..., 80$ should approximate the diagonal. Furthermore, assuming Y has a standard Gumbel distribution and that $Y_1, ..., Y_{80}$ are independent, pointwise confidence bounds can be derived from e.g. Czörgő & Révész (1978).

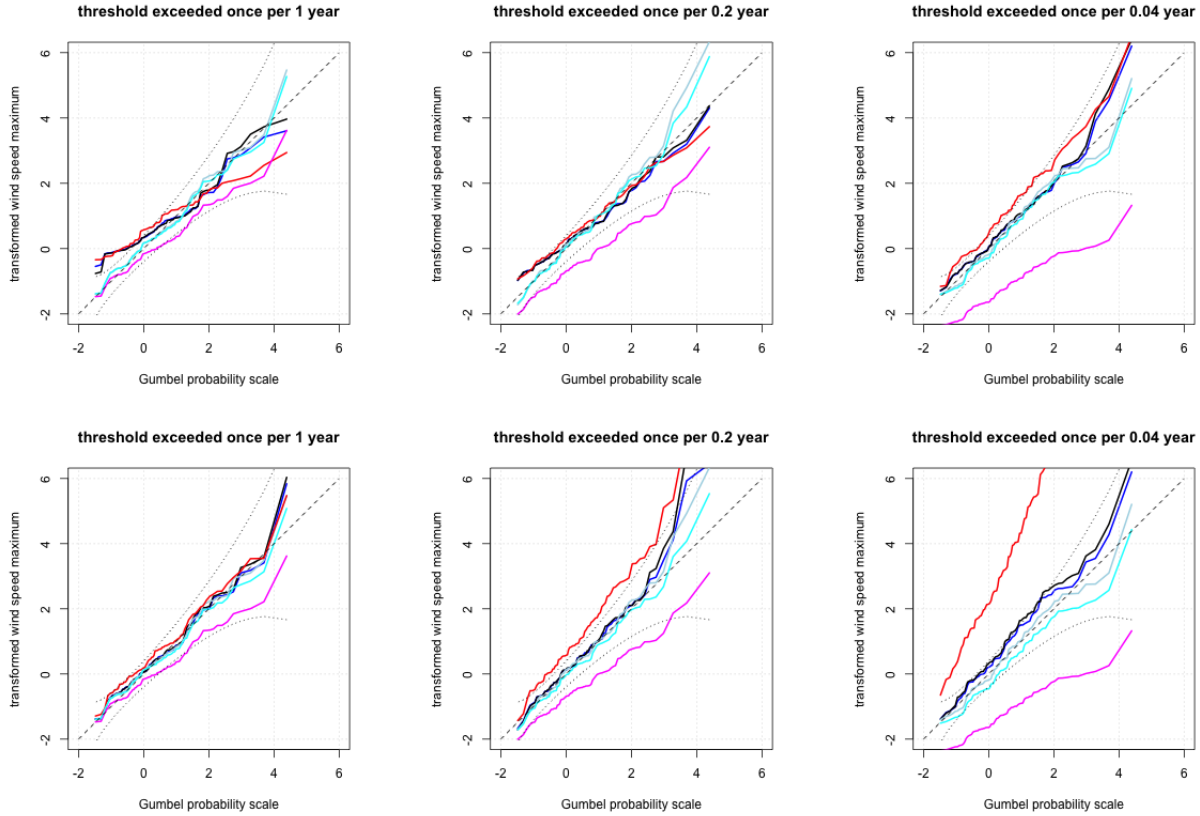

Fig. 1: Gumbel plots of transformed wind speed maxima over 80 subsamples of the System-4 data. Dashed: standard Gumbel line; dotted: 90% confidence bounds. Transformations are based on tail fits: GP (red), exponential (magenta), GW (blue), log-GW (black), GW with shape parameter equal to 0.5 (cyan), and 1-parameter Weibull tail (light blue). Top/bottom: same/different thresholds for scale and shape estimation.

The accuracy of the approximation provides a measure of the quality of the extrapolation: on each subsample, the tail model is fitted on a certain number of wind speed values (the highest l wind speeds, say), but the test only concerns the highest wind speed of each subsample, so we test the extrapolation from a probability $p = l/n$ (with n the size of the subsample) to a probability $1/n$.

The deviation of $Y_{i:80}$ against $-\log(-\log(i/81))$ from the diagonal is a measure of deviation from the classical tail limit for maxima (which is equivalent to the GP tail limit). As it does not go beyond this classical limit, a small deviation from the diagonal for a particular tail model does not guarantee that this model is suitable for extrapolation over orders of magnitude in frequency.

Figure 1 shows the plots of $Y_{i:80}$ against $-\log(-\log(i/81))$ with $Y_1, ..., Y_{80}$ estimated using $l = 72$, 360 and 1800 wind speed values, which correspond to thresholds exceeded during fractions of time p of 0.00068, 0.0034 and 0.017, respectively, or exceeded with return periods of 1, 0.2 and 0.04 year.

The top panels show results of tail fits using the same thresholds as above for shape parameter estimation (see the end of Section 2). Overall, the curves for the fitted GW, log-GW and 1-parameter Weibull tails approximate the diagonal quite closely. Also shown is a curve for a GW model shape parameter ρ fixed to 0.5, a

value matching reasonably well to estimates of ρ over a range of thresholds. It is close to the other curves. The same applies to the GP tail. However, the curves for the exponential tail deviate much further from the 1:1 line with increasing F; the exponential tail is apparently too heavy to approximate the tail of the System-4 wind speed.

The bottom panels show results of tail fits using higher thresholds for shape parameter estimation. The results agree largely with those in the top panels, with the exception of curves for the GP tail, which now deviate further from the 1:1 line for the lower return periods. The larger error is likely due to bias in the shape parameter estimates at these lower thresholds.

For the 1-parameter Weibull tail, these findings are in line with the results of a similar outlier check on local Weibull fits to ERA-40 reanalysis wind speed data in van den Brink and Können (2008) and van den Brink and Können (2011). As the 1-parameter Weibull tail is a special case of the log-GW tail and furthermore, equivalent to a GW tail with positive shape parameter, it is not surprising that the results for the latter tail models are similar to those for the 1-parameter Weibull tail.

3.3 Extrapolation of wind speed return values from subsamples

A second check compares estimates of the return value x_R for a large return period R from the subsamples with a reference estimate of x_R obtained from the entire sample, with both estimates based on the same tail model.

For R, we could choose a return period smaller than the length of time covered by the full dataset, so well below 5000 years, to ensure that the reference estimates are accurate. But this would have the same limitation as the check on maxima over subsamples in Section 3.2: it remains well within the scope of the classical extreme value limit (i.e., (7) with the ratio on the left-hand side bounded away from 0), whereas what we really want to know is how the different models of the tail perform in extrapolation over several orders of magnitude in return period, which goes beyond the classical tail limit. Therefore, we focus on large return periods; we report the case $R = 10^7$ year, which is an upper bound to the range of return periods we are interested in. Results for smaller R, e.g. $R = 1000$ years are similar to those for $R = 10^7$ year, mainly differing in the absolute magnitudes of the deviations between estimates from subsamples and estimates from the full sample.

Estimates of x_R for $R = 10^7$ year were computed from the 80 subsamples of size $n = 105192$ (each covering 72 years). These were derived from the tail approximations (12)-(15) with all or most parameters estimated from the highest l values in the subsample[3], with l ranging from 10 to 6000 and with sample fraction $p = l/n$ in (12)-(15) ranging from about 10^{-4} to about 0.05. These estimates were compared to reference estimates of x_R derived from the entire dataset using the same tail approximation; so estimates of x_R from the subsamples based on the exponential tail, for example, are compared to reference estimates of x_R from the full sample which

[3] Only for the GW and log-GW models, the shape parameter is estimated from a larger number of values; see Appendix B.

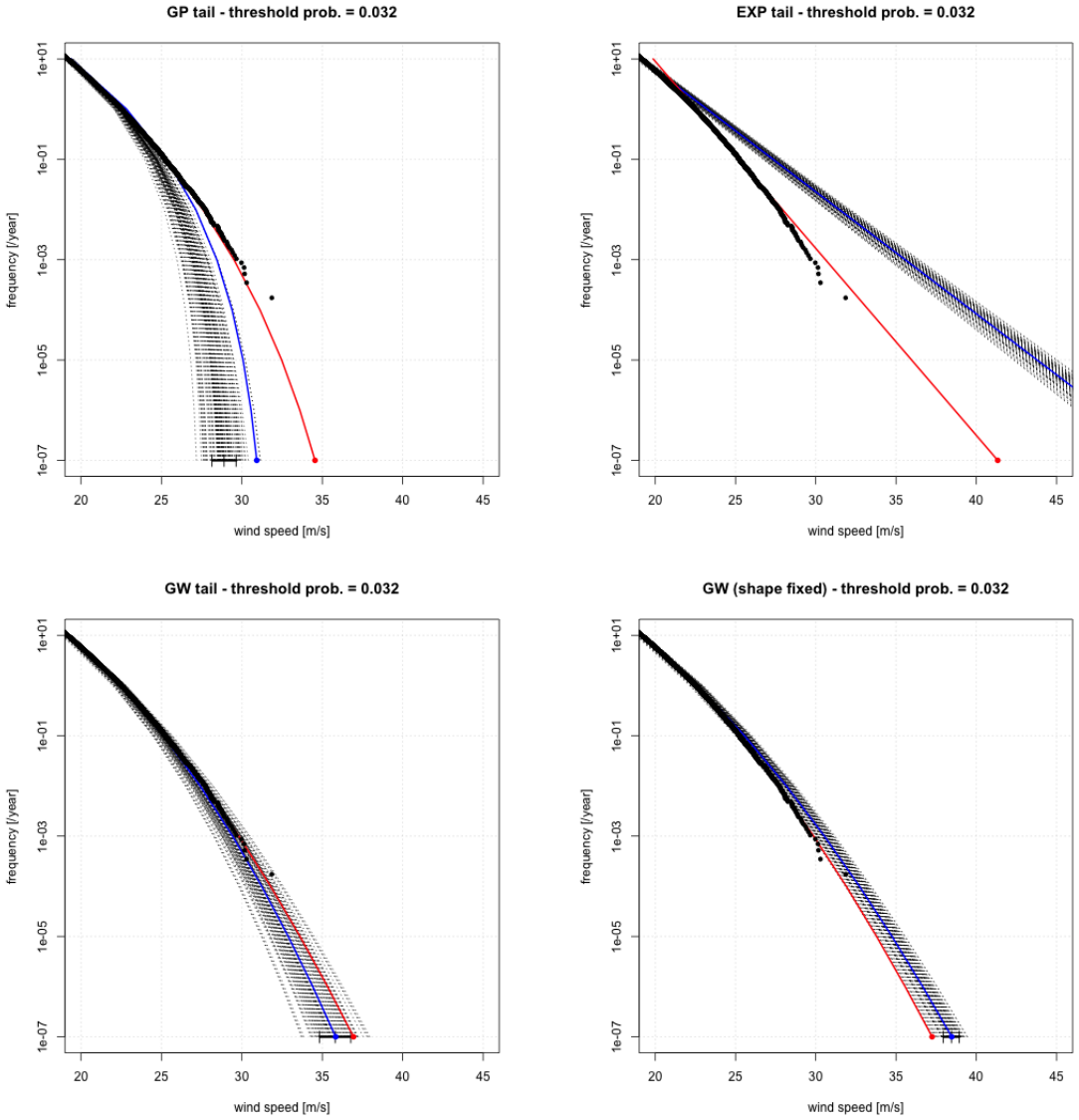

Fig. 2: Tails estimated from subsamples (black dotted curves) from the 3.2% highest values in each subsample ($p = 0.032$) of the System-4 data for four tail models (top left to bottom right): GP, exponential, GW and GW with fixed shape parameter $\rho = 0.5$. Mean estimate of of x_R with error bar at a return period of $R = 10^7$ years in black. Black dots: order statistics of the full sample. Blue: tail estimated from the $p = 0.032$ highest values of the full sample. Red: same, for $p = 0.032/80 = 4 \cdot 10^{-4}$. Estimates made with different thresholds for shape and scale.

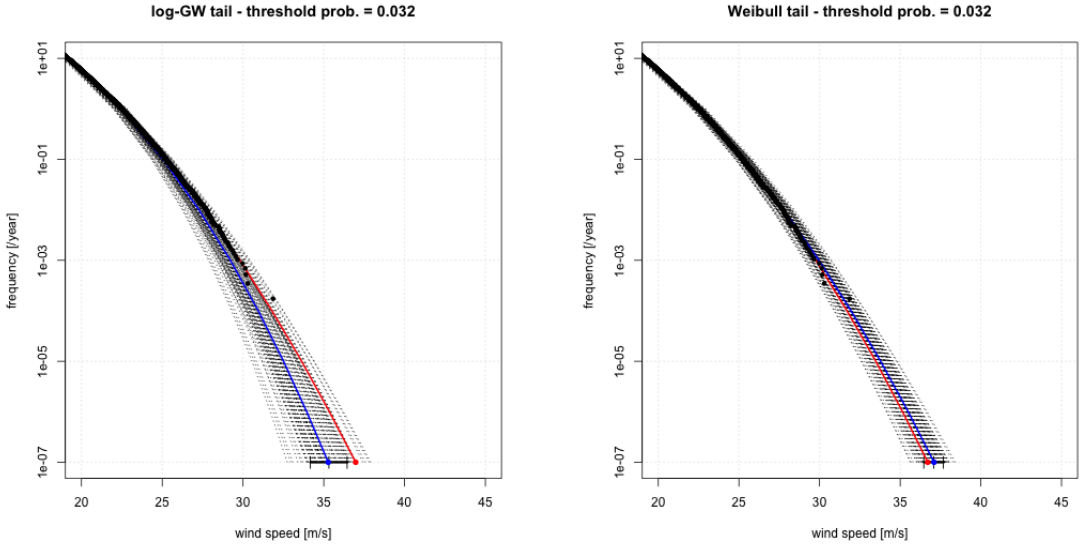

Fig. 2: Continued: same for the log-GW tail (left) and 1-parameter Weibull tail (right).

are also based on the exponential tail. Figure 2 shows the estimates for a single value of the sample fraction p: for each of the GP, exponential, GW, log-GW and 1-parameter Weibull tail models, it shows the tails estimated from the subsamples (dotted lines), and for a frequency of 10^{-7} /year (corresponding to $R = 10^7$ year), it shows the mean $\langle \hat{x}_R \rangle$ of the estimates of x_R from the subsamples and the interval $\langle \hat{x}_R \rangle \pm \sigma_R$, with σ_R the sample standard error of the estimates.

In addition, each panel in Figure 2 shows two reference estimates of the tail from the entire sample. The blue curve is the tail estimated using the same values of the sample fraction p as for the estimates from the subsamples (so n and l are both larger by a factor of 80). Because p is the same in both cases, the subsample-based estimates and the reference estimates should have the same bias, so the results should merely reflect the additional variance of the estimates from the subsamples. Indeed, at a frequency of 10^{-7} /year, the blue curve almost coincides with the mean $\langle x_R \rangle$ of the estimates from the subsamples except for the GP tail (for the latter, estimates from the subsamples are highly variable).

The red curves in Figure 2 represent reference estimates from the complete dataset derived with a value of p which is 80 times smaller than the value used for the estimates from the subsamples (so the value of l used for the reference estimates and the subsample-based estimates is the same). As the ratio p/p_R in (12)-(15) is now a factor of 80 larger for the subsample-based estimates than for the reference estimate, we are really testing how well the different tail models extrapolate[4]. The statistics of the deviations of the dotted (subsample-based) curves from the red curves based on the full sample at a frequency of 10^{-7} year ($R = 10^7$ year) are summarized in Figure 3 for the full range of p. This is the figure we will now describe in detail.

[4] Note that the variability of the red curves is higher than the variability of the blue curves, as the red curves are based on an 80 times smaller number of order statistics.

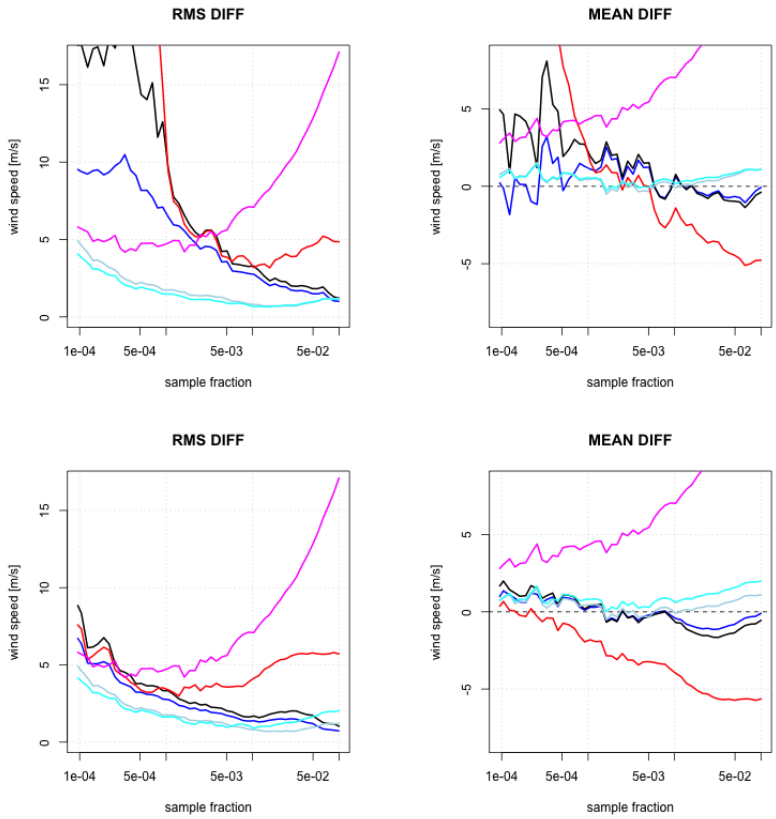

Fig. 3: RMS difference (left) and mean difference (right) between subsample-based estimates of the return value x_R of wind speed and a reference estimate from the entire System-4 dataset as function of the sample fraction p used in the subsample-based estimation. Top/bottom: same/different thresholds for scale and shape. Tail models: GP (red), exponential (magenta), GW (blue), log-GW (black), GW with shape parameter 0.5 (cyan), 1-parameter Weibull (light blue). $R = 10^7$ years.

The top panels show the root-mean square (RMS) and mean difference between the subsample-based estimates of x_R and the reference estimate of x_R from the complete dataset as a function of p, for estimates using the same thresholds for scale and shape. Over the entire range of p, the mean difference is smallest in magnitude for the 1-parameter Weibull tail and the GW tail with fixed shape. It is somewhat larger for the GW and log-GW tails with estimated shape; the values are noisy (due to noise in the shape estimates: they are based on the same threshold as used for scale estimation). The mean difference is much larger in magnitude for the GP tail and in particular, for the exponential tail. These findings are also reflected in the RMS differences on the left-hand side, in particular for higher values of p where the mean difference is generally larger than the standard deviation. For low values of p, the RMS for the exponential tail is lower than for the GW, log-GW and GP tails, as the exponential has a fixed shape parameter, whereas the estimates of the shape of GW, log-GW and GP tails are highly variable, being estimated with the same threshold as used for the scale.

The bottom panels show results obtained with shape estimated with a lower threshold than scale (results for the exponential, 1-parameter Weibull, and GW with fixed shape are identical to those in the top panels, as these models have no

shape parameter to be estimated). The patterns are overall similar, but at low p, estimates based on the GW, log-GW and GP tails are much more precise, and the mean difference for GP is now somewhat larger in magnitude. For the 1-parameter Weibull, GW and log-GW models, the highest sample fraction still gives about the lowest RMS difference, indicating that bias is nowhere dominant for these tail models.

Both the estimates from the full sample and from the subsamples are biased and imprecise. This is why we did not speak of bias and RMS error, but used the terms "mean difference" and "RMS difference" in the discussion of the differences between full-sample and subsample estimates.

However, bias and RMS error of the estimates are what we are really after. Fortunately, there is a way to estimate these from the results just obtained.
We can make a crude order-of-magnitude estimate of the size of the bias in an estimate $\hat{x}_{R,s}$ of x_R from a 72-year subsample. Let $\hat{x}_{R,f}$ be the estimate from the full sample. Using an asymptotic expression of bias in the high quantile estimator, we can approximate the ratio

$$\frac{\mathbb{E}(\hat{x}_{R,s} - x_R)}{\mathbb{E}(\hat{x}_{R,s} - \hat{x}_{R,f})}; \tag{18}$$

the method is described in Appendix A for each of the tail models considered.
The ratio (18) is positive, and depends on the shape parameter of a second-order extension of the tail model. Letting that parameter tend to 0 from below (which corresponds to reducing the speed of convergence to the tail limit) maximizes (18), so this gives a worst-case approximation of this ratio.
The denominator in (18) is estimated by the mean difference displayed in Figure 3 (bottom right). Multiplying it by the worst-case ratio (18) gives a worst-case estimate of $\mathbb{E}(\hat{x}_{R,s} - x_R)$, the bias in the estimate $\hat{x}_{R,s}$ from a subsample.
The standard deviation $\sigma_{R,s}$ of the error in the estimate $\hat{x}_{R,s}$ of x_R from a subsample can be estimated directly from the estimates of $\hat{x}_{R,s}$. To derive the standard deviation $\sigma_{R,f}$ of the error in the estimate of x_R from the full sample, we multiply the estimate of $\sigma_{R,s}$ by an approximation of $\sigma_{R,f}/\sigma_{R,s}$ derived from known large-sample limits. Then the RMS error of quantile estimates from subsamples and from the full dataset can be derived straightforwardly from the estimates of their bias and standard deviation.

The estimates of bias in the return values estimated from the full set of System-4 wind speed data are shown in Figure 4 (blue). It should be stressed that these bias estimates are derived independently for each of the tail models (GP, EXP, GW, log-GW, 1-parameter Weibull). The exponential tail was regarded as a special case of the GW tail in developing the second-order model. The bias estimates in Figure 4 are based on tail estimates using different thresholds for shape and scale (i.e., the results in the bottom panels of Figure 3).

For the 1-parameter Weibull, GW and log-GW tails, the estimated bias is very small. Curiously, the magnitude of this bias estimate for the GW tail with estimated shape parameter is smaller than for the GW tail with fixed shape, even though the

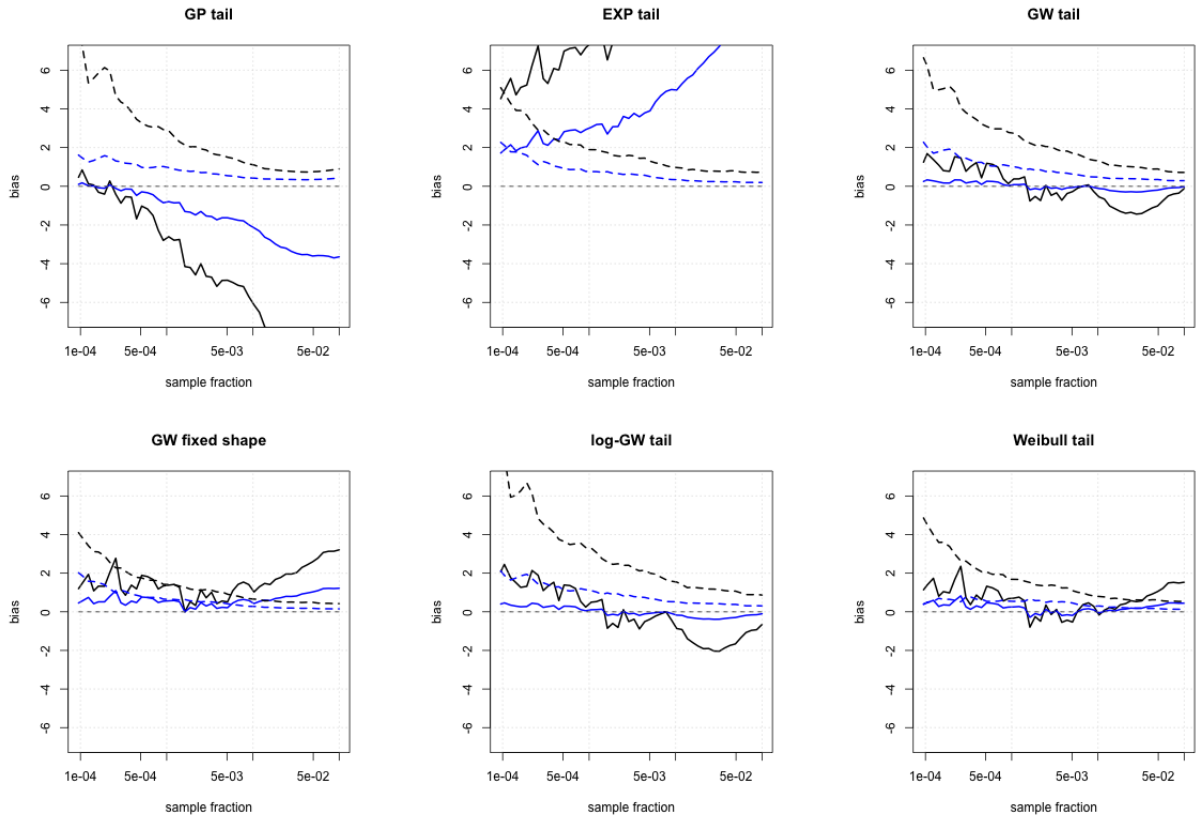

Fig. 4: Estimates of worst-case bias in the estimates of the return value x_R of wind speed from the full System-4 dataset (blue) and from a 72-year subsample (black) for estimates based on the (top left to bottom right) GP tail, exponential tail, GW tail, GW with fixed shape parameter 0.5, log-GW tail and 1-parameter Weibull tail. Dashed lines: standard errors of estimates of the return value (same colour codes). Return period $R = 10^7$ years. Thresholds for shape and scale are different.

mean differences in Figure 3 (bottom right) are similar in magnitude. The reason is that estimation of the shape parameter results in partial cancellation of the bias in estimates of high quantiles; see Appendix A. The estimated bias is much larger for the GP and exponential tails, also in comparison with the standard errors (dashed).

To summarize, disregarding potential systematic errors in the System-4 forecast model and errors due to the 6-hourly sampling, it appears to be feasible to estimate the 10^7 year wind speed accurately with valid confidence bounds by applying a GW, log-GW or 1-parameter Weibull tail fit to the full set of System-4 ensemble forecast wind speed data.

In the same manner, using the bound on (18), we can estimate the worst-case bias in $\hat{x}_{R,s}$ obtained from a subsample; see Figure 4 (black). The values are considerably larger in this case. However, for the GW, log-GW and 1-parameter Weibull tails, the estimated bias remains below 2 m/s over most of the range of p. Similar values can be expected for estimates from the longest records of wind measurements in the Netherlands, as these are of about the same size as the subsamples from the ensemble forecast wind speeds considered here.

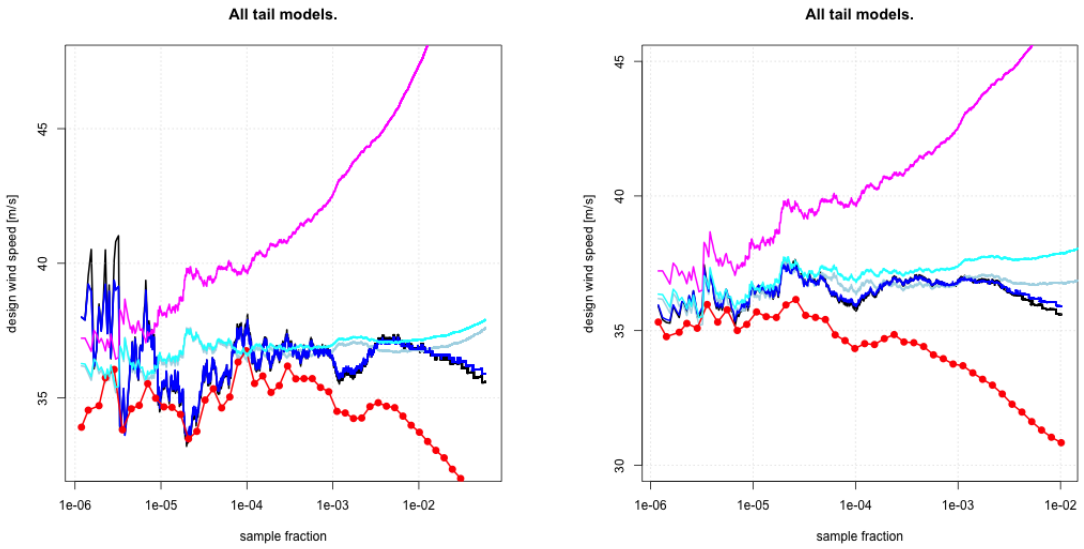

Fig. 5: Estimates of 10^7 year wind speed from the complete System-4 dataset as function of the sample fraction used in the estimation. Tail models: GP (red), exponential (magenta), GW (blue), GW with fixed shape (cyan), log-GW (black), and 1-parameter Weibull (light blue). Left/right: same/different thresholds for scale and shape.

3.4 Estimates of the 10^7 year wind speed from the complete dataset

Estimates of the 10^7 year wind speed from the complete dataset are shown in Figure 5. The estimates based on the GW tail with fixed shape and the 1-parameter Weibull tail are very stable as a function of sample fraction p. Those based on the GW and log-GW models tend to increase somewhat with decreasing p (i.e., with increasing wind speed threshold). The estimates based on the GP tail vary more with p and are lower. For the exponential tail, the estimates are higher, and show a steep trend as a function of p; this model is clearly not suitable for the System-4 data.

The estimates based on different thresholds for shape and scale are displayed with confidence intervals in Figure 6. For the exponential tail, the confidence interval is evidently not valid; the large trend in the estimates as a function of p indicates that bias dominates the error.

3.5 Consistency of return values of water level and wind speed

A potential application of the large seasonal ensemble forecast datasets of ECMWF is the estimation of return values of high-tide water level along the coast of the Netherlands for very large return periods (van den Brink, 2018). One way to do this is to simulate water levels directly from the wind and pressure forecast data, and then estimate the tails of the local water level simulations. Another approach proposed in Caires et al. (2016) is to estimate the local extreme value statistics of the wind over an area covering the North Sea from the seasonal ensemble forecast data, and use these statistics to scale up the wind fields (and possibly pressure fields) to generate a database representing more extreme conditions then present in

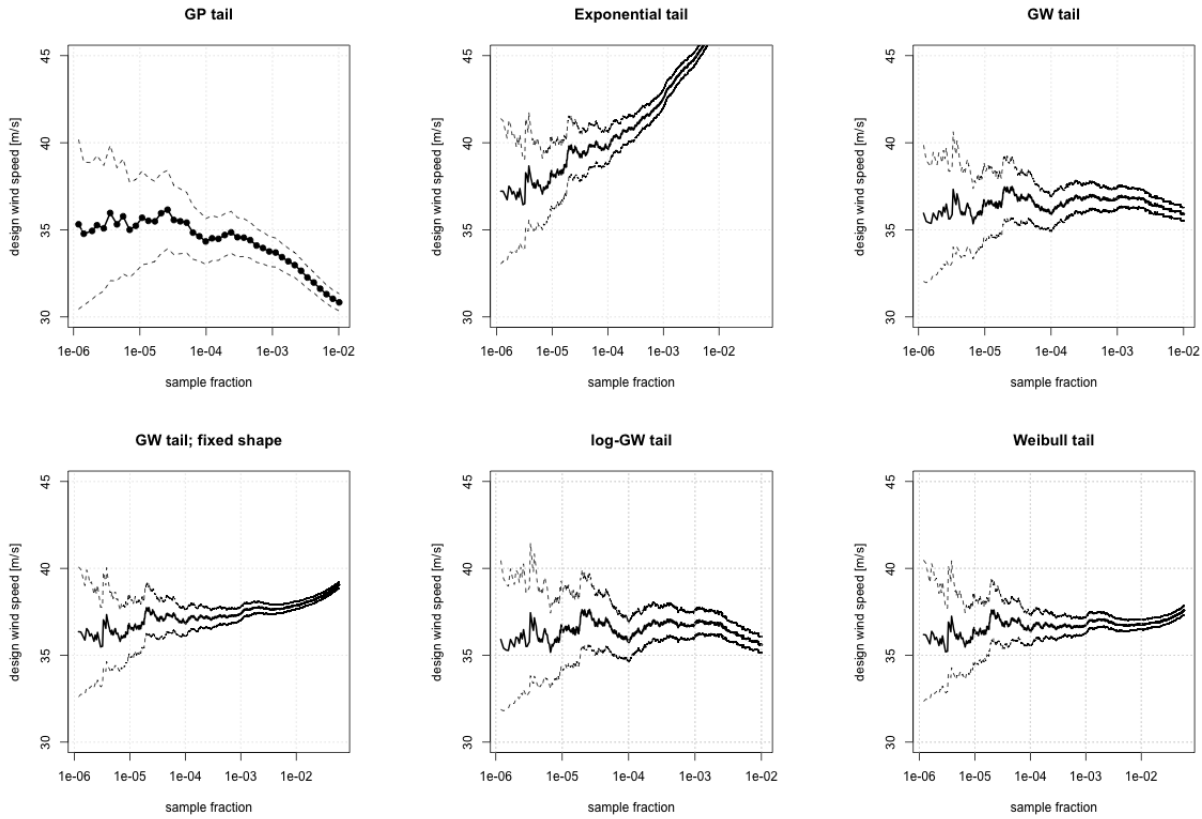

Fig. 6: Same as Figure 5 with 90% confidence bounds: Tails (top left to bottom right): GP, exponential, GW, GW with fixed shape $\rho = 0.8$, log-GW, 1-parameter Weibull. Estimates using different thresholds for scale and shape.

the original data. Subsequently, these can be used to simulate extreme water levels directly.

Ideally, both approaches would produce approximately the same extreme water level statistics. Whether this is so may depend on the model chosen to extrapolate the local tails of the distribution functions of wind speed and/or water level. The effect of this choice can be assessed within a simplified setting, using a simple approximation of the relationship between wind speed and high-tide water level. We used the following equilibrium relationship between high-tide water level ζ and near-surface wind component from north-west u_{NW}:

$$\zeta = c(u_{NW}^2/g) + \zeta_a \tag{19}$$

with g the gravitational acceleration (9.81 m/s^2), ζ_a the astronomical high water level, and c a dimensionless constant. We took $c = 0.06$ and $\zeta_a = 1.1$ m as in De Valk and Zitman (1999). These values are somewhat arbitrary; therefore, the water level values obtained from this model are not necessarily realistic. For the present purpose, this is no problem; it is sufficient that the wind-water level relationship is qualitatively similar to more accurate models.

Figure 7 shows the high-tide water levels derived from the estimates of the 10^7 year north-westerly wind speed by (19) in black, and the 10^7 year high-tide water level estimated from high-tide water level data derived from data of north-westerly

wind speed by (19) in blue. Ideally, these curves are identical. For the GW tail fits, they are almost identical. For the log-GW and 1-parameter Weibull tail fits, they differ much more at the higher sample fractions. The same is observed for the GP tail fits, which are much more variable and uncertain. The difference between the two types of estimates is by far the largest for the exponential tail.

This result agrees with expectation. Applying the exponential tail to both wind speed and water level is inconsistent with (19). The somewhat more flexible GP model can adapt better to the the data than the exponential model, as it allows for different values of the shape parameter for water level and wind speed.

The exponential tail is a special case of the GP tail, namely a GP tail with shape parameter $\gamma = 0$ (see (7)). However, that the estimates of GP tails for wind speed and simulated high-tide water level appear to be more consistent with the model (19) than exponential tail estimates, should not be seen as evidence that the true values of the GP shape parameters of wind speed and simulated water level are unequal.
In fact, it can be show that when $\gamma \leq 0$ for the tail of wind speed (which it is, as the estimated tails are lighter than an exponential tail), then the tail of water level simulated by (19) from wind speed must have the same value of γ as the tail of wind speed.
There is a better explanation why GP tail estimates appears to be more consistent with (19) than exponential tail estimates: convergence to the proper (common) GP tail limit of wind speed and simulated water level is incomplete, and GP tails with different values of γ for wind speed and water level may (for a given dataset of fixed size) provide a better approximation of the tails than GP tails with the common true value of γ, even though the functional relationship between wind speed and water level inferred from the fitted GP tails does not match (19) exactly.
For example, if the NW wind speed has a Weibull tail with shape parameter $\rho = 1/2$ (i.e., $F(x) = 1 - \exp(-(x/a)^2)$ for some $a > 0$), then the water level has an exponential tail according to (19). The latter is a GP tail with $\gamma = 0$; the former has an exponential tail as its classical GP tail limit, so $\gamma = 0$ for both. However, a Weibull tail can be closely approximated over a fairly wide range of probabilities by a GP tail with a well chosen value of γ, which can be derived analytically from the Weibull shape parameter; this was shown in van den Brink and Können (2011) and Furrer and Katz (2008). Applying such a GP approximation to the Weibull tail of wind speed, we end up with closely matching GP approximations of the tails of both water level and wind speed, which should also match (19) reasonably well.

If $\rho > 0$ as in the estimates, then the GW tail limit (9) implies a Weibull tail limit, so (14) reduces to (16) with $g(p) = \rho$; this formula remains valid (with a different value of ρ) after a power transformation. The same applies to the log-GW tail (as the 1-parameter Weibull tail is a special case of it). So these three tails should be able to accommodate the first term of the transformation (19). The GW tail is invariant under a shift, so it should also accommodate the second term ζ_a accounting for the astronomical tide in (19) as well. In contrast, the 1-parameter Weibull and log-GW tails are not shift-invariant, which likely explains the larger effect of the

transformation on the estimates based on these two tails. If this is correct, then these two tails are more suitable for an analysis of surge level, represented by (19) with $\zeta_a = 0$.

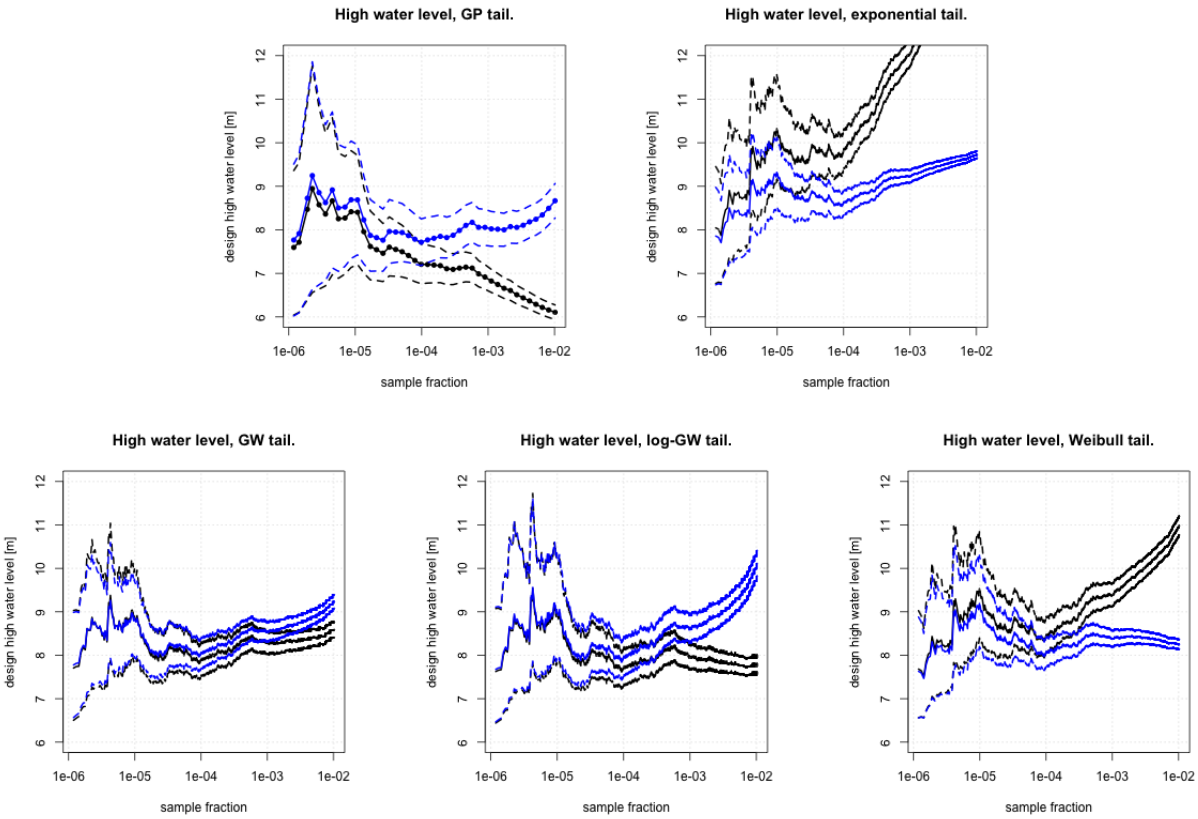

Fig. 7: Estimates of 10^7 year high-tide water level from fitted GP (top left), exponential (top right), GW (bottom left), log-GW (bottom centre) and 1-parameter Weibull (bottom right) tails derived from System-4 data. Black: tails of NW wind component fitted and transformed by (19) to water level. Blue: water level tails fitted on data obtained using (19). Thresholds for shape and scale estimation are different.

4 Analysis of SEAS5 data

4.1 The SEAS5 dataset

The archive of SEAS5 data is similar to that of System-4: seasonal reforecasts have been computed starting at the first day of every month in 1981-2016, each forecast run over at least 7 months. The ensemble size is 25. In addition, the operational forecasts are available for 2017; from these, we used 25 ensemble members of each forecast. The time step is a 6 hours. Further information about the SEAS5 data can be found in Implementation of Seasonal Forecast SEAS5 and SEAS5 User Guide.

Data of wind at 10m height at the location 55N, 3E were retrieved from the SEAS5 archive. Subsamples of the data were generated by combining ensemble members with the same label and with starting dates at regular 6-month intervals. This resulted in 150 subsamples covering 37 years each. Pairs of these were com-

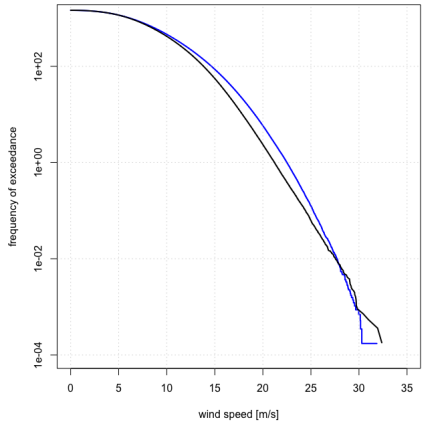

Fig. 8: Empirical frequencies of exceedance of wind speed from System-4 (blue) and SEAS5 (black).

bined, resulting in 75 subsamples, each covering 74 years; approximately the same volume as for System-4.

A comparison of the empirical tails of the distribution functions of System-4 and SEAS5 wind speeds is shown in Figure 8. Each frequency of exceedance in this plot is derived from the fraction of time of exceedance $1 - F$ by multiplying it by αn; see eq. (3).

For SEAS5, the tail is overall lower, but its plot has lower curvature than the tail plot for System-4. As a result, the curves cross at a wind speed of about 28 m/s, above which SEAS5 wind speeds are higher. This discrepancy is worrying: at least one of the two datasets gives biased wind speeds, not just at very high wind speeds but also in the lower range. We will discuss the effects of this difference on the extrapolation of the tail in the following sections.

4.2 Statistics of the highest wind speed in a subsample

The results of the check of tail models based on the maxima of wind speed over the subsamples (see Section 3.2) are shown in Figure 9. As in the case of System-4 data, a GW model with a fixed shape parameter ρ was included among the models to be tested; in this case, a value of 0.8 was chosen, based on estimates from the full set of SEAS5 data. Results are similar to those obtained with System-4 data, except that:

- the exponential tail is performing better than on System-4 data at relatively low thresholds;
- a GW model with fixed GW shape parameter of 0.8 is performing at least as good or better than all other models.
- the 1-parameter Weibull tail performs poorly on SEAS5 data, where its performance on System-4 data is very good.

The first point is the result of the SEAS5 tail being heavier than the System-4 tail (see Figure (8)); however, the second point shows that the exponential tail (which

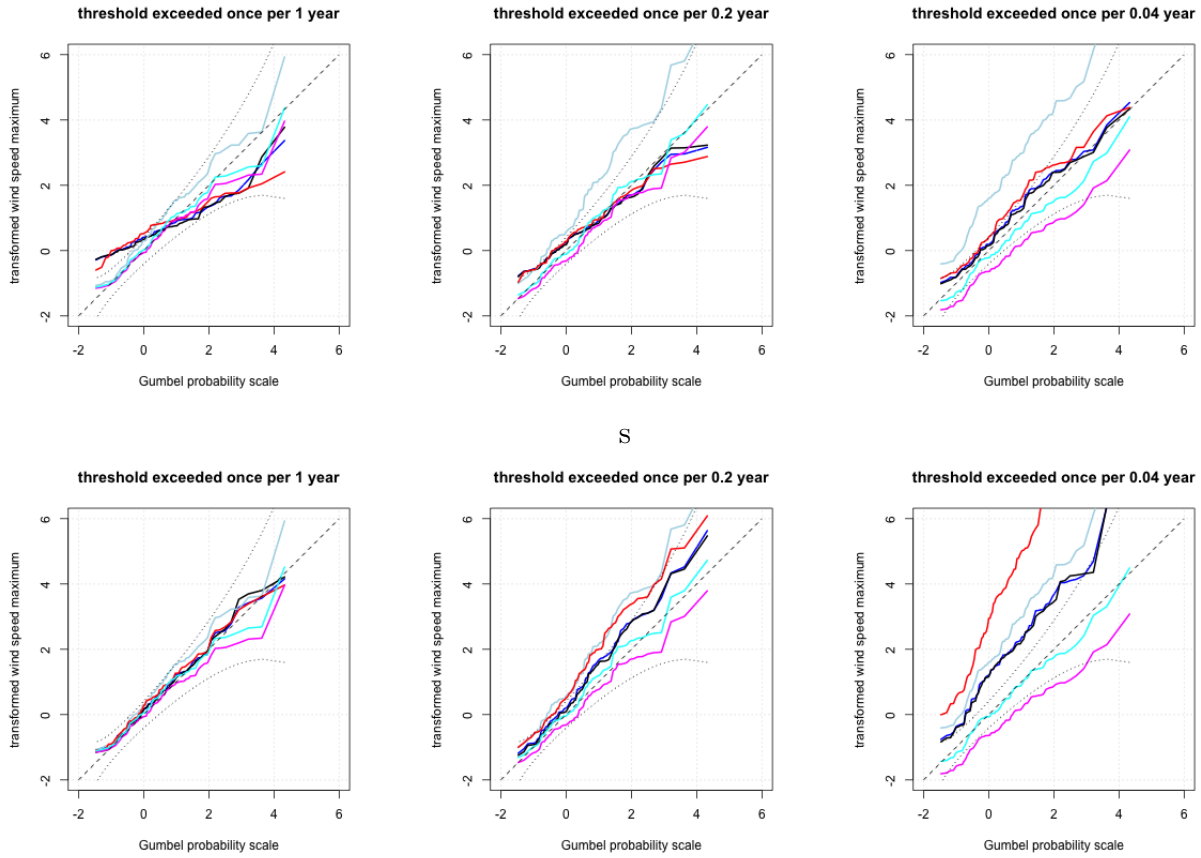

Fig. 9: Gumbel plots of transformed wind speed maxima over 80 subsamples of the SEAS5 data. Dashed: standard Gumbel line; dotted: 90% confidence bounds. Transformations are based on tail fits: GP (red), exponential (magenta), GW (blue), log-GW (black), GW with shape parameter equal to 0.8 (cyan), and 1-parameter Weibull tail (light blue). Top/bottom: same/different thresholds for scale and shape estimation.

is a GW model with GW shape parameter equal to 1) is not necessarily the model of choice in this case; better models exist among the GW tails.

Comparing results for GW, log-GW and GP tails based on the same thresholds for shape and scale estimation (top) with results based on a lower threshold for shape (bottom), the latter are clearly worse than the former. This is most likely due to lack of regularity of the tail of the SEAS5 data; see Section 4.4.

Again, results indicate that the classical GP tail limit applies with $\gamma = 0$ (the exponential tail), since for the higher thresholds, all models perform satisfactorily, including the exponential tail.

4.3 Extrapolation from subsamples

Referring to Section 3.3 for the background of the analysis of extrapolation from subsamples, we proceed with a discussion of differences between the results from the SEAS5 and System-4 datasets.

Figure 10 shows differences between return value estimates from subsamples and reference estimates from the entire dataset based on sample fractions 75 times lower than for the estimates from subsamples. Compared to Figure 3 for System-4 data, we observe the following differences.

- For the exponential tail, the magnitude of the mean difference is reduced considerably: this is to be expected, as the tail of SEAS5 data is heavier.
- For SEAS5 data, RMS differences are larger than for System-4 data for all tail models except the exponential. The same holds for the magnitude of the mean difference. This indicates again that the SEAS5 tail is less regular than the System-4 tail.
- The GP tail shows the largest differences (both in terms of RMS and mean) of all tail models.
- A GW tail with fixed shape parameter equal to 0.8 has consistently the smallest mean difference and RMS of all tail models. This implies that the exponential tail (a GW tail with shape parameter equal to 1) is not the best model for the SEAS5 tail. However, the more flexible tails with estimated shape parameter (GP, GW and log-GW tails) are apparently sensitive to irregularity of the tail.

Extrapolations of these differences to estimates of worst-case bias in the estimates of return values (see Section 3.3) are shown in Figure 11.

- For the GP tail, estimated bias is very large.
- For the exponential tail, estimated bias is comparable in magnitude to the bias for the GW and log-GW tails over a large range of sample fractions. For estimates from the full sample (blue curves), bias remains largely below 2 m/s for these models. Bias for the 1-parameter Weibull tail is somewhat larger.
- For the GW tail with fixed shape parameter equal to 0.8, bias is much lower than for the other tails, including the GW tail with estimated shape; it even

remains below 2 m/s for estimates from subsamples (black curves) over a wide range of sample fractions. Apparently, estimation of the shape parameter contributes considerably to the bias in this case, due to irregularity of the SEAS5 tail (see also Section 6).

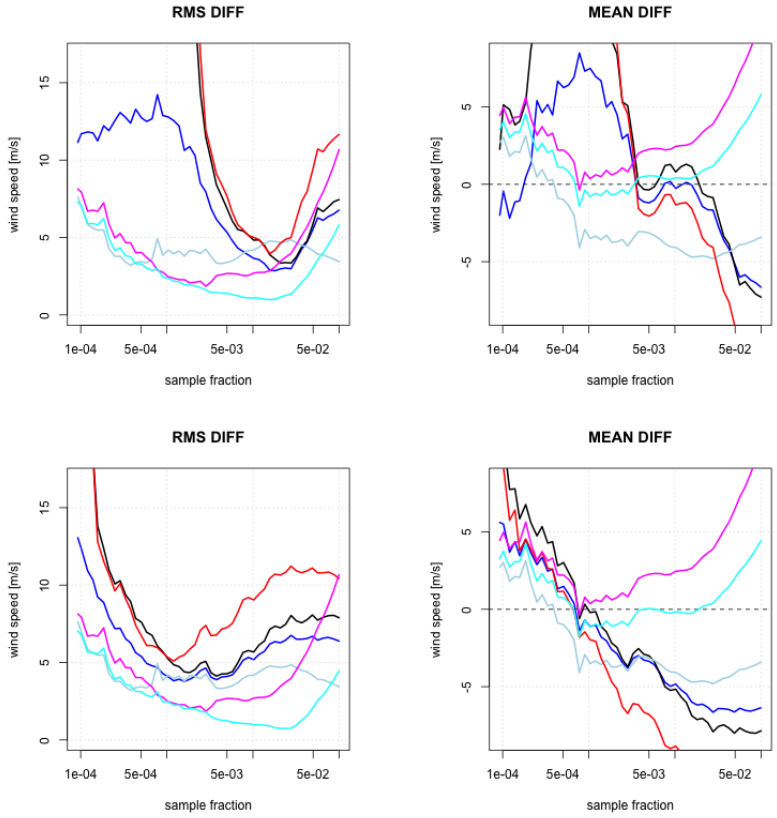

Fig. 10: RMS difference (left) and mean difference (right) between subsample-based estimates of the return value x_R of wind speed and a reference estimate from the entire SEAS5 dataset as function of the sample fraction p used in the subsample-based estimation. Upper/lower: same/different thresholds for scale and shape. Tail models: GP (red), exponential (magenta), GW (blue), log-GW (black), GW with shape parameter 0.8 (cyan), 1-parameter Weibull (light blue). $R = 10^7$ years.

4.4 Estimates of the 10^7 year wind speed from the complete dataset

For all tail models, the estimates of return values from the full set of SEAS5 data vary considerably as a function of the sample fraction; see Figure 12. This shows again that the tail of the full SEAS5 dataset is irregular; compare the much more stable estimates from System-4 data in Figure 5. It explains the rather large bias for almost all tail models found in the previous section: all models assume some notion of regularity of the tail.

Estimates based on different thresholds for shape and scale are generally much more stable as functions of sample fraction; for sample fractions in the range 10^{-5}

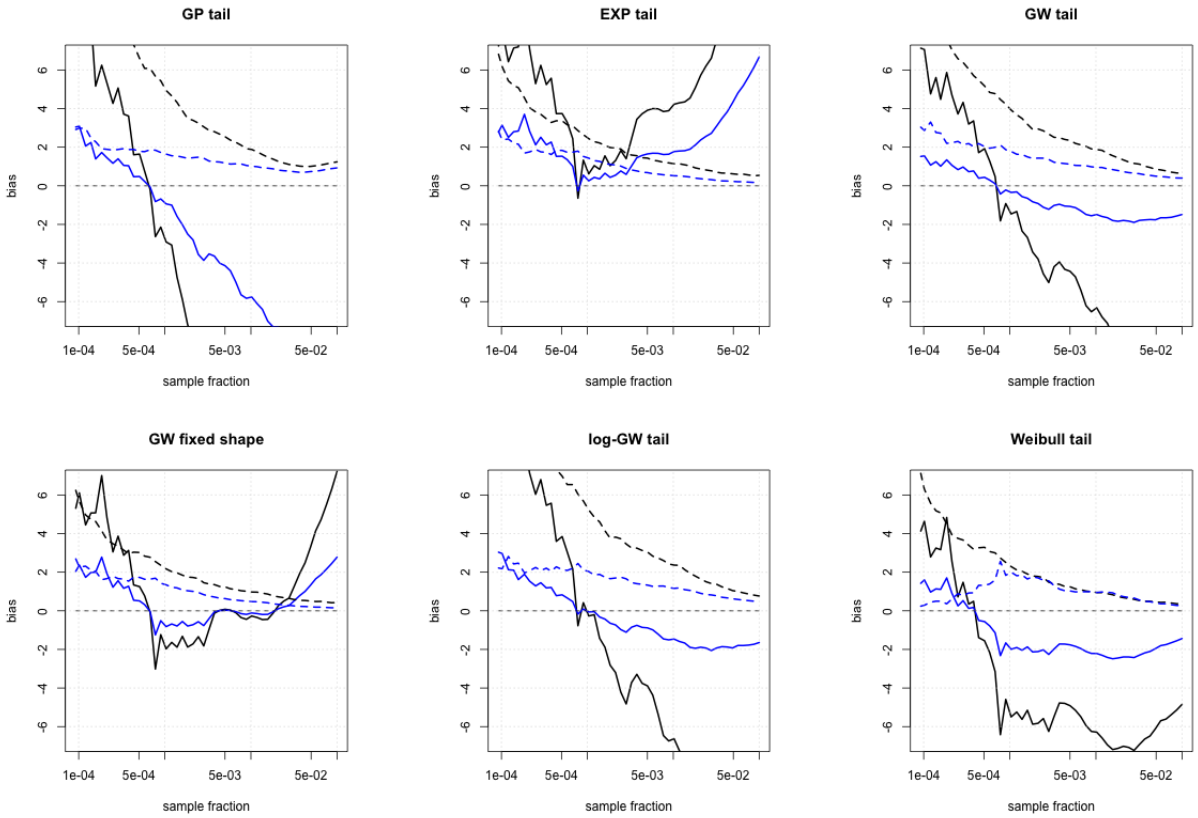

Fig. 11: Estimates of worst-case bias in the estimates of the return value x_R of wind speed from the full SEAS5 dataset (blue) and from a 72-year subsample (black) for estimates based on the (top left to bottom right) GP tail, exponential tail, GW tail, GW tail with fixed shape parameter $\rho = 0.8$, log-GW tail and 1-parameter Weibull tail. Dashed lines: standard errors of estimates of the return value (same colour codes). Return period $R = 10^7$ years.

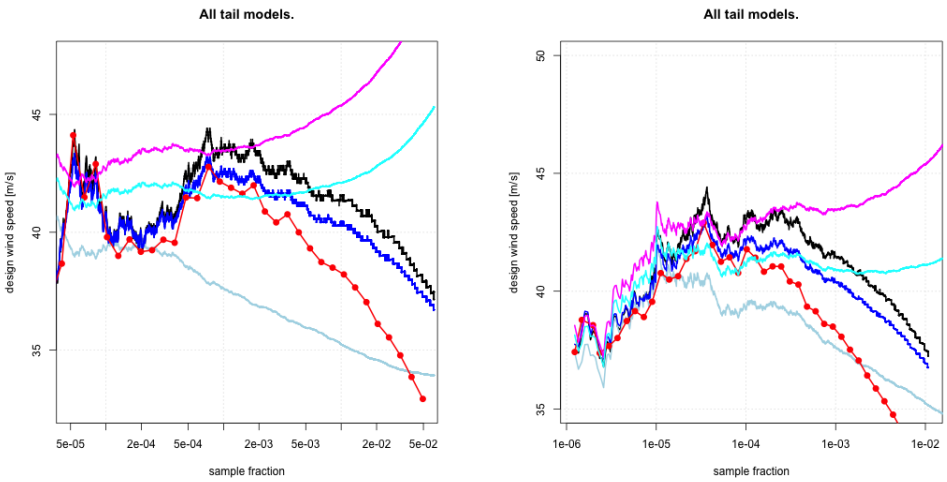

Fig. 12: Estimates of 10^7 year wind speed from the complete SEAS5 dataset. Tails: GP (red), exponential (magenta), GW (blue), GW with fixed shape $\rho = 0.8$ (cyan), log-GW (black) and 1-parameter Weibull (light blue). Left/right: same/different thresholds for scale and shape.

to $5 \cdot 10^{-4}$, the estimates based on the GW, log-GW, exponential and GP tails are all fairly stable and not very different from each other.

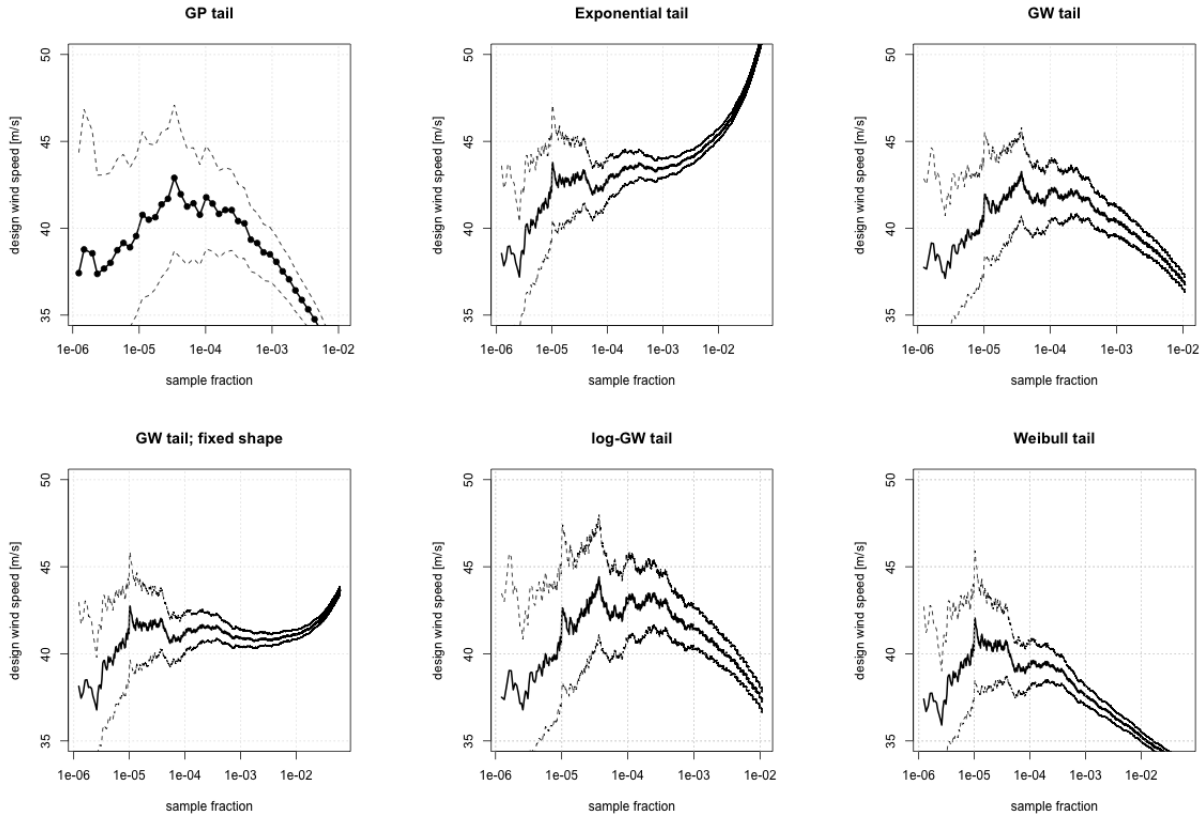

Fig. 13: As Figure 12, with 90% confidence bounds (dashed). Tails (top left to bottom right): GP, exponential, GW, GW with fixed shape $\rho = 0.8$, log-GW, 1-parameter Weibull. Estimates using different thresholds for scale and shape.

In line with Figure 8, the estimates of the return value of wind speed from the full set of SEAS5 data are much higher than the estimates from the full set of System-4 data; compare Figure 12 to Figure 5. This is not only a matter of scale: the GW shape parameter estimates are much higher for SEAS5 than for System-4 (not shown).

4.5 Consistency of return values of water level and wind speed

The results from SEAS5 are similar to those obtained from System-4 data; compare Figure 7 to Figure 14.

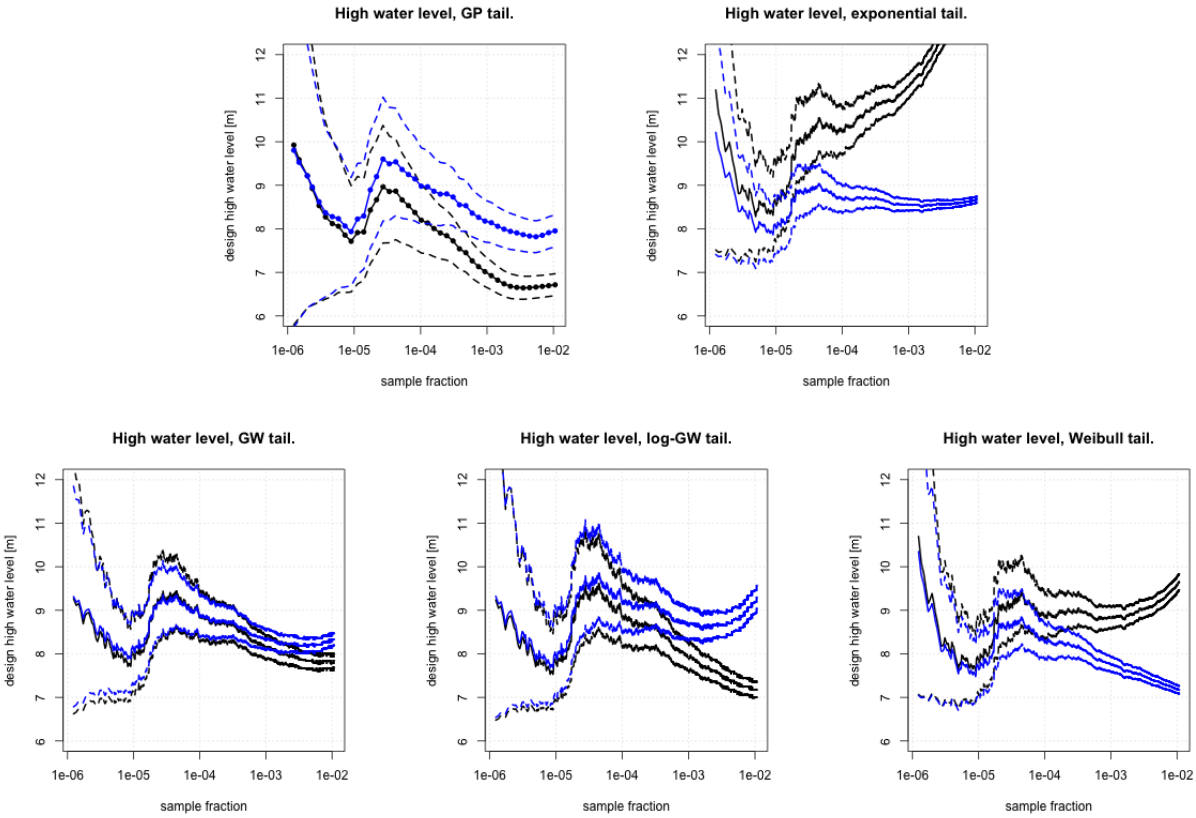

Fig. 14: Estimates of 10^7 year high-tide water level from fitted GP (top left), exponential (top right), GW (bottom left), log-GW (bottom centre) and 1-parameter Weibull (bottom right) tails derived from SEAS5 data. Black: tails of NW wind component fitted and transformed by (19) to water level. Blue: water level tails fitted on data obtained using (19). Thresholds for shape and scale estimation are different.

5 Analysis of Speedy annual wind speed maxima

5.1 The Speedy dataset

Speedy is a low-resolution climate model covering the globe; see Molteni (2003); Kucharski et al (2006). A total of 12816 Speedy runs were made covering the years 1980-2009 with observed sea surface temperatures and initialized at rest, using a spin-up over 1 years. Only the annual maxima of near-surface wind speed were saved; the total number of wind speed maxima is 384480. We considered data at the same location as for the analysis of seasonal forecast data.

These Speedy data were used to test the performance of estimation of the 10^7 year wind speed from a dataset covering approximately 5700 years, approximating the lengths of the System-4 and SEAS5 dataset. To this end, 67 subsamples of the Speedy wind speed maxima were extracted form the data-set, each subsample covering 5700 years (similar to the total length of the System-4 and SEAS5 data). On these subsamples, the same checks were performed as previously applied to the System-4 and SEAS5 data. A difference with the latter checks is that from the Speedy data, we only used the annual maxima.

We fitted the same tail models as before: GW, log-GW, 1-parameter Weibull, exponential and GP. As an alternative, we could have used tail models specifically aimed for use on block-maxima, e.g. the Generalized Extreme Value (GEV) instead of the GP tail, the Gumbel instead of the exponential, etc. However, when estimating from a reasonably small fraction of the data (say from values above the 100-year return value), it does not matter much whether such an adjustment for block-maxima is made. This is illustrated by Figure 15, showing a Gumbel tail (of let's say the distribution function of annual maxima) and its approximation by an exponential tail.

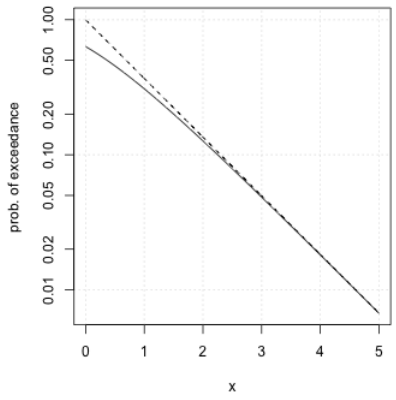

Fig. 15: Gumbel tail (full) and its exponential approximation (dashed).

5.2 Statistics of the highest wind speed in a subsample

The results of the comparison of tail models based on the maxima of wind speed over the subsamples (see Section 3.2) are shown in Figure 16. The pattern is not

very different from the results for System-4 and SEAS5 data, but for Speedy data, the GW and log-GW tails appear to fit particularly well in comparison to the 1-parameter Weibull, exponential and GP tails.

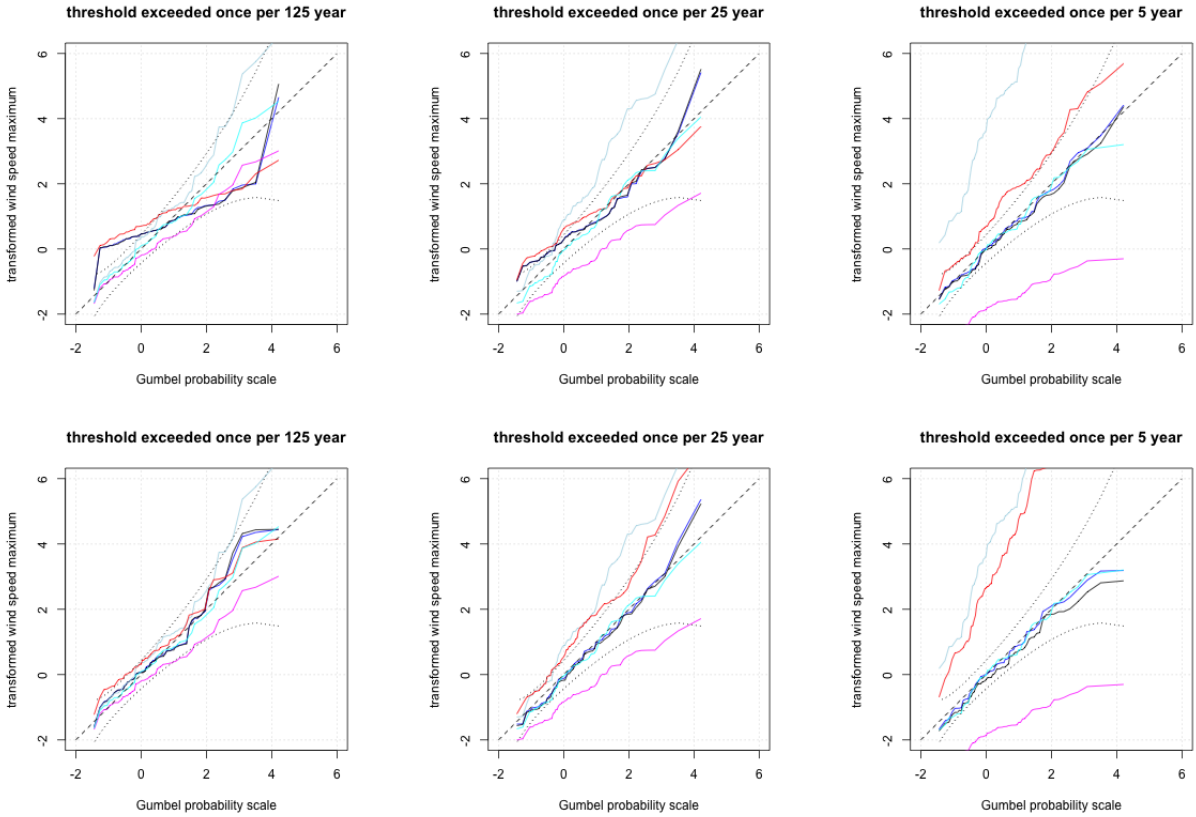

Fig. 16: Gumbel plots of transformed wind speed maxima over 67 subsamples of the Speedy data. Dashed: standard Gumbel line; dotted: 90% confidence bounds. Transformations are based on tail fits: GP (red), exponential (magenta), GW (blue), log-GW (black), GW with shape parameter equal to 0.5 (cyan), and 1-parameter Weibull tail (light blue). Top/bottom: same/different thresholds for scale and shape estimation.

5.3 Extrapolation from subsamples

Both the top and bottom panels of Figure 17 show that random error is much smaller than with System-4 or SEAS5 data (Figures 3 and 10); note the different vertical scales of the plots. This is expected, as subsamples from the Speedy data cover 5700 years, to be compared with 72 and 74 years for subsamples from the System-4 and SEAS5 data. For estimates for very high sample fractions close to 1, the mean difference is large for all models, which is likely the result of using tail models not specifically adapted to block maxima (see Section 5.1).

- For the exponential tail, the mean difference is very large; this is related to the rather light tail of the Speedy data. For the GP and 1-parameter Weibull tails, large values of the mean difference are obtained.

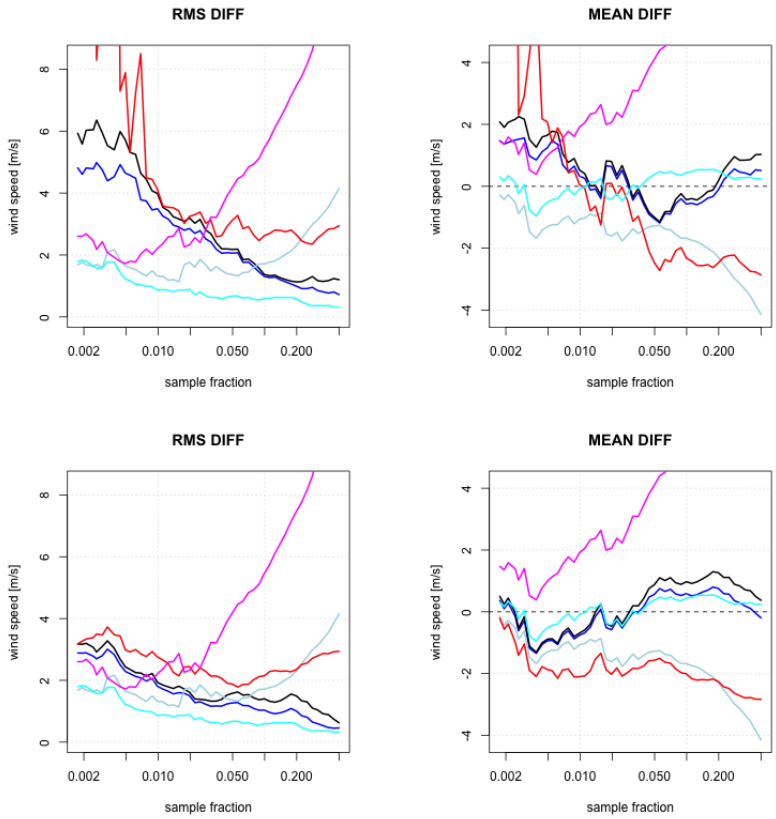

Fig. 17: RMS difference (left) and mean difference (right) between subsample-based estimates of the return value x_R of wind speed and a reference estimate from the entire Speedy dataset as function of the sample fraction p used in the subsample-based estimation. Upper/lower: same/different thresholds for scale and shape. Tail models: GP (red), exponential (magenta), GW (blue), log-GW (black), GW with shape parameter 0.5 (cyan), 1-parameter Weibull (light blue). $R = 10^7$ years.

- For the GW and log-GW tails with fitted shape parameter, the mean difference is small and the RMS is low for the larger sample fractions: for these, it makes little difference whether the shape parameter is estimated, or fixed at the value of 0.5. The decrease of RMS with sample fraction observed over almost the entire range of sample fractions indicates that bias is low for this models.

Extrapolations of these results to estimates of worst-case bias in the estimates of return values are shown in Figure 18. In this case, we need to focus on the black lines representing worst-case bias from 5700 year subsamples. Note again that vertical scales are different from those in the plots for System-4 and SEAS5 data.

- For the GP and 1-parameter Weibull tails, bias is large, and of similar order of magnitude as the large standard error.

- For the exponential tail, bias is very large, in particular when compared to the low standard error. Confidence bounds based on the standard error are not valid for this model.

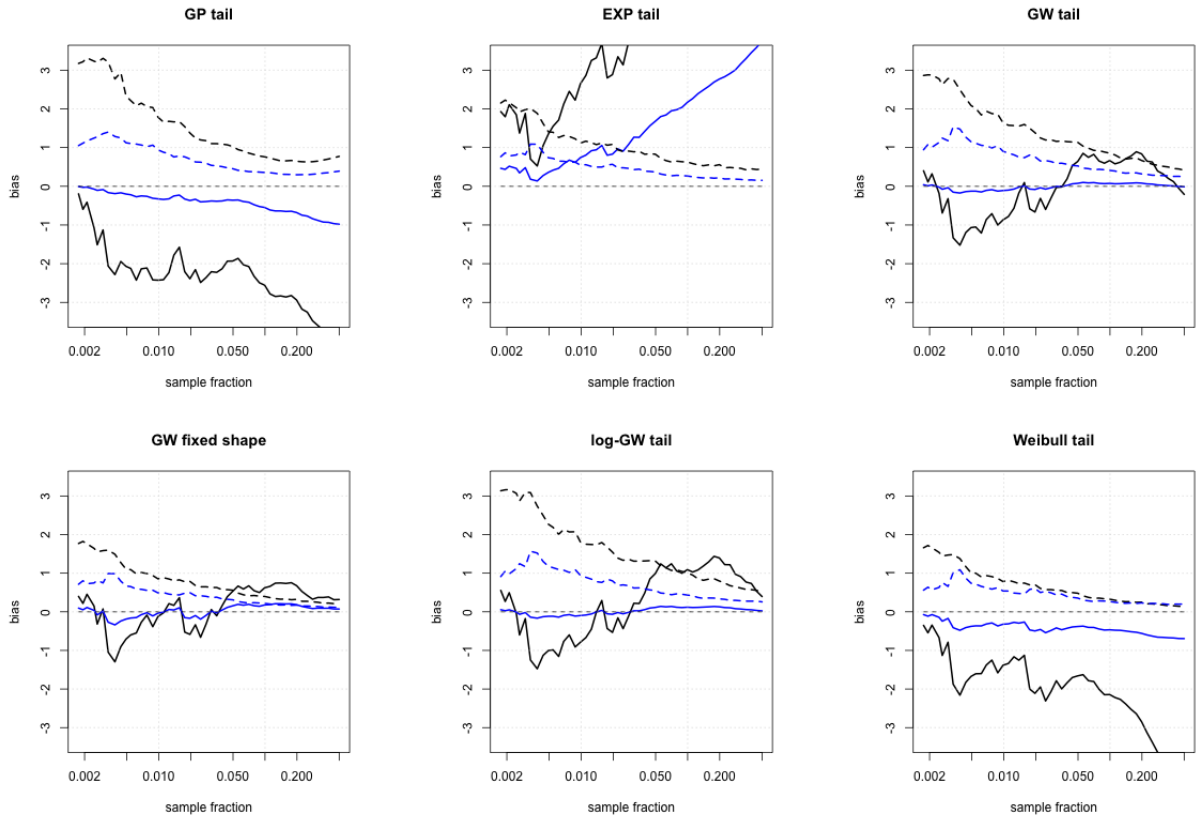

Fig. 18: Estimates of worst-case bias in the estimates of the return value x_R of wind speed from the full Speedy dataset (blue) and from a 5700-year subsample (black) for estimates based on the (top left to bottom right) GP tail, exponential tail, GW tail, GW tail with fixed shape parameter $\rho = 0.5$, log-GW tail and 1-parameter Weibull tail. Dashed lines: standard errors of estimates of the return value (same colour codes). Return period $R = 10^7$ years.

- For the GW and log-GW tails, bias is small: it remains mostly below 1 m/s. Also the ratio of bias to standard error remains acceptable for most sample fractions, in particular for the GW tail.

5.4 Estimates of the 10^7 year wind speed from the complete dataset

Of all four tail models, the GW and log-GW tails produce the most stable estimates of the return value of wind speed from the full set of Speedy annual wind speed maxima; see Figure 19. Estimates based on the GP and 1-parameter Weibull tails are more variable, but mostly in the same range. For the exponential tail, the estimates vary strongly with threshold.

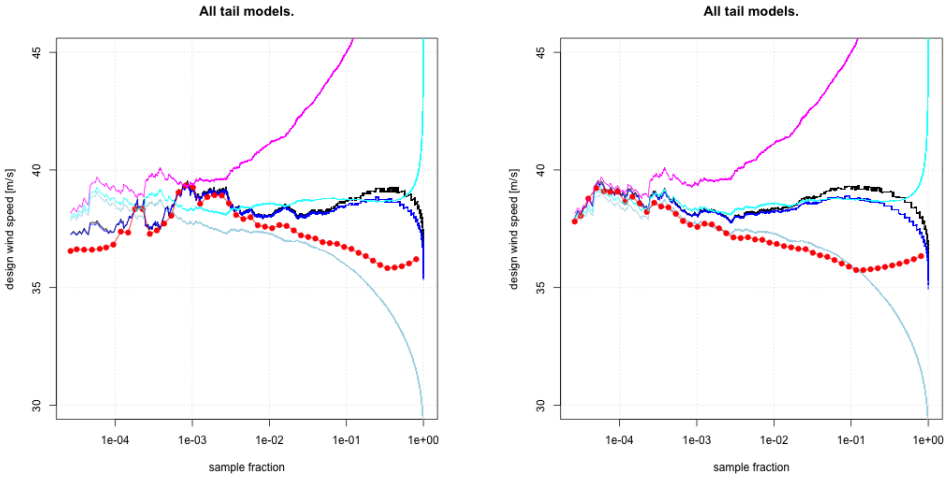

Fig. 19: Estimates of 10^7 year wind speed from the complete Speedy dataset. Tails: GP (red), exponential (magenta), GW (blue), GW with fixed shape $\rho = 0.5$ (cyan), log-GW (black) and 1-parameter Weibull (light blue). Left/right: same/different thresholds for scale and shape.

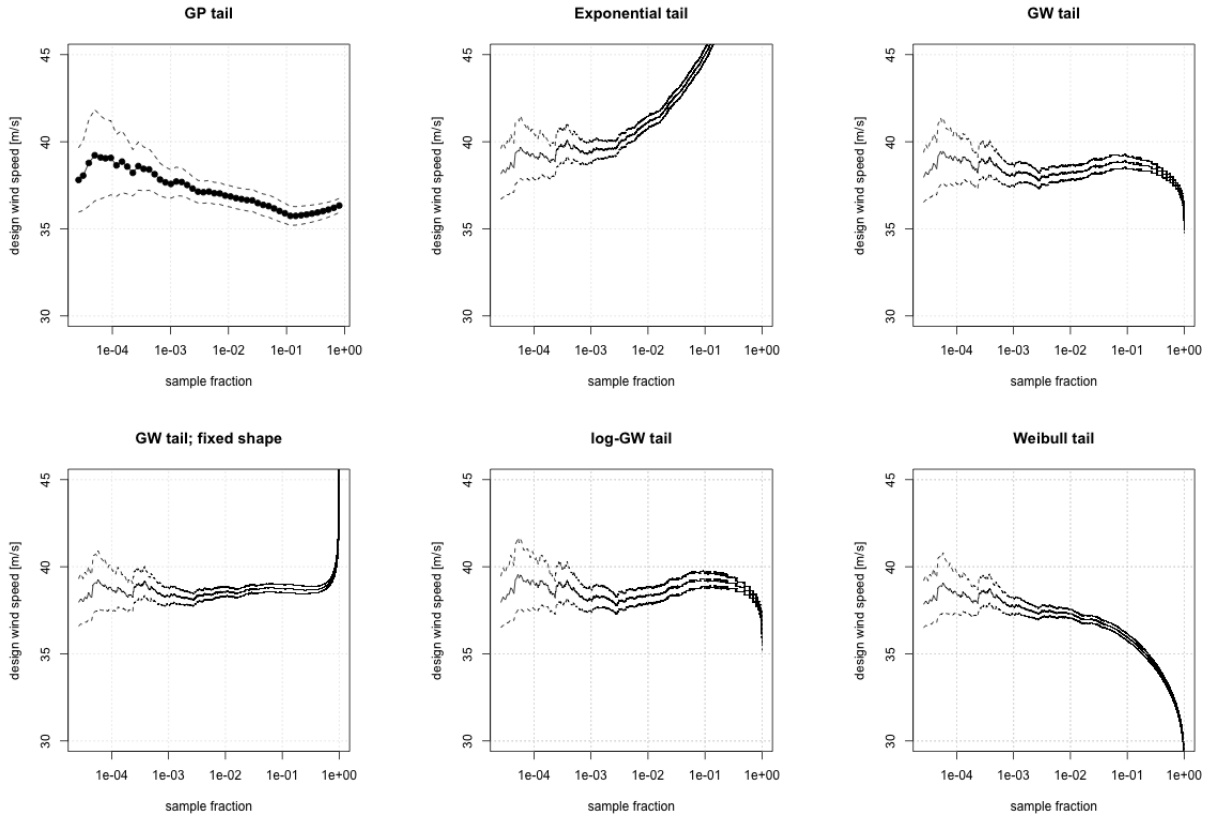

Fig. 20: As Figure 19, with 90% confidence bounds (dashed). Tails (top left to bottom right): GP, exponential, GW, GW with fixed shape $\rho = 0.5$, log-GW, 1-parameter Weibull. Estimates using different thresholds for scale and shape.

6 Discussion

Consequences of tail model and sample size The present comparison of estimates of wind speed with a very large return period obtained using different models of the tail of the distribution function of wind speed provides valuable insight about the impact of the size of the available dataset and of the choice of model of the tail.

Tables 1-4 summarize median values of worst-case bias (see Section 3.3 and Appendix A) and standard deviation of estimates of the 10^7 year wind speed over selected ranges of thresholds, with sample fractions p as indicated[5]. Note that for Speedy, the dataset covers a much longer period of time and consists of annual maxima, so the p range has an entirely different meaning than for the other datasets. For estimates based on different thresholds for shape and scale (Tables 3 and 4), lower sample fractions p were used than for estimates based on identical thresholds for shape and scale (Tables 1 and 2), because the smallest errors tend occur at lower p.

A drawback of such summaries is that the error magnitudes listed are not the best achievable magnitudes for an individual data-set and tail. Therefore, we discuss general tendencies rather than individual numbers.

Table 4 shows that for estimation of the 10^7 year wind speed, really accurate estimates with RMS error of around 1 m/s require an unusually large dataset such as provided by the ECMWF seasonal ensemble forecast archives. With SEAS5, this may not be achievable. With approximately 70 years of data, an RMS error of about 2-3 m/s appears to be feasible with the best performing models on System-4 data. With SEAS5 data, having a less regular tail, this may not be feasible due to bias.

Overall, the best performing model is the GW (Generalized Weibull) tail. Estimates based on the GP (Generalized Pareto) model have consistently higher bias, and are almost all cases also more variable. Estimates based on an exponential tail are strongly biased on two of the three datasets; evidently, this one-parameter tail model lacks flexibility. As a result, confidence intervals based on standard errors of the exponential tail fit are not reliable. It is not unlikely that the same applies to the confidence intervals for wind speed in WBI-2017; see Chbab, 2017). The 1-parameter Weibull tail suffers from a similar lack of flexibility, but it performs considerably better than the exponential tail on two out of three datasets; the RMS values in tables 2 and 4 are quite good. Estimates with the GW tail are never much worse than with other tails; for two of the three datasets, they are among the best. Furthermore, GW tail fits preserve a typical wind-surge relationship much better than other tails.

The choice of a model of the tail cannot solve all problems in extrapolation: the tail of the SEAS5 data lacks regularity (see e.g. Figure 12), leading to considerable bias in estimates for all tail models except the GW tail with shape parameter fixed to 0.8, a crude "eyeball" estimate from the full sample of SEAS5 data. The latter exception is good news for the use of datasets of the size of the SEAS5 data, since a reasonable guess of the shape can be made based on these data. However, if much shorter records of measurement data are used and the tail is as irregular as the

[5] The medians are taken over thresholds exceeded by sample fractions p equidistant on a logarithmic scale.

Dataset	length [yr]	p range	GP	EXP	GW	GW ρ fixed	log-GW	Wbl
System-4	72	0.005-0.05	4.3(4.3)	14(1.7)	0.6(4.2)	1.5(1.1)	0.7(5.3)	0.4(1.5)
SEAS5	74	0.005-0.05	3.3(7.5)	4.5(1.6)	1.3(5.4)	0.4(1.3)	1.4(7.8)	6.6(1.4)
Speedy	5700	0.05-0.5	3.3(3.0)	9.3(1.8)	0.6(2.9)	0.6(1.1)	0.6(3.5)	2.5(1.0)
System-4	5760	$(0.7\text{-}7.0)\cdot10^{-4}$	1.6(1.7)	5.8(0.5)	0.1(1.7)	0.6(0.4)	0.1(2.0)	0.1(0.5)
SEAS5	5550	$(0.7\text{-}7.0)\cdot10^{-4}$	1.3(2.5)	1.9(0.6)	0.3(2.3)	0.2(0.5)	0.3(2.8)	2.2(0.6)
Speedy	381900	$(0.7\text{-}7.0)\cdot10^{-3}$	0.8(0.9)	2.6(0.3)	0.1(0.9)	0.2(0.2)	0.1(0.9)	0.5(0.2)

Tab. 1: Worst-case bias and standard deviation (in brackets) of estimates of the 10^{-7} year wind speed from (sub)samples of three datasets of different lengths, for five different tail models: values are medians over the indicated range of sample fraction p, in m/s. Estimates using identical thresholds for scale and shape.

Dataset	length [yr]	p range	GP	EXP	GW	GW ρ fixed	log-GW	Wbl
System-4	72	0.005-0.05	6.8(33)	14(49)	4.2(37)	2.0(38)	5.3(36)	1.6(37)
SEAS5	74	0.005-0.05	9.5(37)	4.8(46)	6.0(40)	1.5(41)	8.1(41)	6.7(35)
Speedy	5700	0.05-0.5	4.3(36)	9.5(46)	3.0(39)	1.3(39)	3.5(39)	2.7(35)
System-4	5760	$(0.7\text{-}7.0)\cdot10^{-4}$	2.7(36)	5.8(41)	1.7(37)	0.8(37)	2.0(37)	0.5(37)
SEAS5	5550	$(0.7\text{-}7.0)\cdot10^{-4}$	3.6(40)	2.0(43)	2.4(40)	0.7(41)	2.8(41)	2.3(39)
Speedy	381900	$(0.7\text{-}7.0)\cdot10^{-3}$	1.2(39)	2.6(40)	0.9(39)	0.3(38)	0.9(39)	0.6(37)

Tab. 2: Worst-case root-mean square error and mean value (in brackets) of estimates of the 10^{-7} year wind speed from (sub)samples of three datasets of different lengths, for five different tail models: values are medians over the indicated range of sample fraction p, in m/s. Estimates using identical thresholds for scale and shape.

Dataset	length [yr]	p range	GP	EXP	GW	GW ρ fixed	log-GW	Wbl
System-4	72	0.001-0.01	4.7(3.6)	8.7(2.2)	0.4(2.6)	1.1(1.6)	0.5(3.3)	0.5(2.0)
SEAS5	74	0.001-0.01	9.8(5.8)	2.1(2.2)	4.3(3.3)	1.1(1.9)	3.7(4.6)	5.5(2.1)
Speedy	5700	0.01-0.1	2.2(3.0)	4.4(1.6)	0.6(2.0)	0.4(1.1)	0.5(2.3)	1.8(1.0)
System-4	5760	$(1.4\text{-}14)\cdot 10^{-5}$	1.6(1.2)	3.6(0.8)	0.1(0.9)	0.4(0.7)	0.1(1.0)	0.2(0.8)
SEAS5	5550	$(1.4\text{-}14)\cdot 10^{-5}$	3.6(2.3)	0.9(1.1)	1.0(1.4)	0.5(1.0)	0.9(1.9)	2.0(1.2)
Speedy	381900	$(1.5\text{-}15)\cdot 10^{-4}$	0.4(0.7)	1.3(0.5)	0.1(0.5)	0.1(0.4)	0.1(0.5)	0.4(0.4)

Tab. 3: Worst-case bias and standard deviation (in brackets) of estimates of the 10^{-7} year wind speed from (sub)samples of three datasets of different lengths, for five different tail models: values are medians over the indicated range of sample fraction p, in m/s. Estimates using different thresholds for scale and shape.

Dataset	length [yr]	p-range	GP	EXP	GW	GW ρ fixed	log-GW	Wbl
System-4	72	0.001-0.01	5.8(32)	8.9(45)	2.7(37)	1.9(38)	3.3(36)	2.0(37)
SEAS5	74	0.001-0.01	11(36)	3.1(44)	5.3(39)	2.3(41)	6.0(40)	6.0(36)
Speedy	5700	0.01-0.1	3.7(36)	4.6(43)	2.1(39)	1.3(39)	2.4(39)	2.1(37)
System-4	5760	$(1.4\text{-}14)\cdot 10^{-5}$	1.9(35)	3.7(40)	0.9(37)	0.8(37)	1.0(37)	0.8(37)
SEAS5	5550	$(1.4\text{-}14)\cdot 10^{-5}$	4.3(41)	1.7(43)	1.8(42)	1.1(41)	2.1(43)	2.3(40)
Speedy	381900	$(1.5\text{-}15)\cdot 10^{-4}$	0.8(38)	1.4(40)	0.5(38)	0.4(38)	0.5(38)	0.6(38)

Tab. 4: Worst-case root-mean square error and mean value (in brackets) of estimates of the 10^{-7} year wind speed from (sub)samples of three datasets of different lengths, for five different tail models: values are medians over the indicated range of sample fraction p, in m/s. Estimates using different thresholds for scale and shape.

SEAS5 tail, poor estimates are expected. Therefore, verification of the regularity of the tail is essential, in particular when relatively short data records are used to make estimates for high return periods.

In the latter case, verification of regularity is inherently more difficult than if a long data record is available. However, it may be done implicitly by applying a proper method (well-supported by theory) for choosing a threshold for estimation of the shape parameter: in the case of low regularity, a high threshold (small sample fraction) is chosen, and the variance of the estimator will be high. By this mechanism, low regularity is automatically reflected in high estimates of uncertainty.

Applicability of GW tail For wind, GW-based estimates of return values are generally lower than estimates based on the exponential tail, which are positively biased for the datasets analyzed on the present study. The variance is generally

higher for a GW tail than for an exponential or 1-parameter Weibull tail, but if the shape is estimated at a lower threshold than the scale, then the variance is not much higher.

For the GW, log-GW tail and GP tails estimated with a lower threshold for shape than for scale, the threshold for shape was determined automatically from the threshold for scale (see Section B). This is likely to contribute to the relatively large bias in GW-based estimates of the 10^7 year wind speed from SEAS5 data: the tail of the SEAS5 is irregular, so a low threshold for the shape (exceeded by a large fraction of the sample) leads to large bias.

In such a case, choosing the threshold for the shape based on a plot of the shape estimate as function of threshold (before estimating the scale parameter and return values) would be better than automatic selection of the threshold. A method exists for threshold selection based on the sizes of fluctuations in estimates, using theoretical results from de Valk & Cai (2018) and an approach similar to Drees & Kauffman (1998) and Boucheron & Thomas (2015). The consistently low bias estimates for GW tails with fixed shape indicate that this approach could give better results than obtained on this study: for each data-set, the fixed shape was in fact chosen after inspecting plots of the threshold-dependence of the shape estimates and of the values of a fluctuation size metric.

The results of the present study indicate that the GW tail could provide more reliable confidence intervals for estimates of return values of wind speed than those currently employed in flood safety assessment (Chbab, 2017). The return value estimates presently used (see Chbab (2017), Section 5.4) are based on exponential tail fits, i.e., a GP tail (7) with $\gamma = 0$ (or equivalently, a GW tail (9) with $\rho = 1$). The present study shows that confidence intervals based on this model are likely too narrow, and not realistic. Because the exponential tail tends to give a positive bias in return values, the underestimation of uncertainty may not lead to underestimation of risk. Fitting a GW tail instead of an exponential tail is expected to produce less conservative estimates with somewhat wider confidence intervals (both more realistic); an analysis of wind measurements may show which effect is more dominant.

The GW tail may also be applicable to other random load variables. For wave parameters, there is already much experience with fitting of a conditional 2-parameter Weibull distribution to peaks over thresholds, starting with De Ronde et al (1995); the GW tail generalizes this and provides a rigorous framework for statistical inference, with better estimators of return values and their uncertainty, and methods to assess regularity of the tail and to select a threshold.

For coastal high water level, benefits of the log-GW tail were demonstrated in de Valk & Cai (2018) in a small case study comparing log-GW to GP estimates (the latter implemented as the VVM-0 method presented in Dillingh et al (1993)). In particular, confidence intervals and fluctuations in estimates were found to be much smaller for the log-GW tail. Because the tails of high water level along the Dutch coast are close to exponential tails, the GW tail would most likely perform at least as well as the log-GW tail. The same may apply to shallow inland waters if wind has a large impact on the water level.

For precipitation in the Netherlands, GW and log-GW tails appear to provide

accurate approximations of the tails of rainfall for durations from 10 min up to 9 days. This would extend to river discharge, if not for anomalies in the river discharge tail due to the water level-dependent water storage capacity upstream. The GRADE model (Hegnauer et al, 2014) represents these anomalies explicitly, so it is expected to produce better results than the fit of a regular tail. Possibly, a GW tail might be useful to extrapolate to very large return periods for which GRADE estimates are not available, if needed.

Applicability of the 1-parameter Weibull tail and the exponential tail The 1-parameter Weibull tail is less flexible than the (two-parameter) GW tail, but the performance in terms of RMS in tables 2 and 4 is quite good, due to the low variance of return value estimates. Its weakness is bias; but bias can be detected, e.g. from the threshold dependence of estimates of its parameter and of return values. Therefore, if estimates based on the 1-parameter Weibull tail are stable as a function of threshold, this tail may be preferred over the more variable estimates based on the GW tail. However, one should be cautious about the use of confidence intervals for return values based on the 1-parameter Weibull tail, as variability tends to be smaller than bias for this tail.

These remarks apply also to the exponential tail, being a one-parameter model. However, in the present study, stability of parameter or return value estimates based on the exponential tail was not observed.

Temporal variability An important issue not discussed in this report is the temporal variability of estimates. Preliminary results of an analysis of data of potential wind speed at Schiphol (not yet reported) indicate that variability on long (multi-year) time scales affects all tail quantiles by the same time-varying factor. This factor appears to scale with the root-mean square wind speed. Variability on these long time scales (e.g. Franzke et al, 2015) is not accounted for in current estimates of return values of wind speed and their confidence intervals in Chbab (2017), but it can be a major source of uncertainty. It is different in seasonal forecast data such as System-4 and SEAS5 than in measurements, if only because the seasonal forecasts run for less than a year. It would be valuable to develop a reliable method to assess uncertainty due to variability on long time-scales, and to properly account for it in estimates derived from measurements as well as in estimates derived from seasonal forecast data.

Seasonal forecast and climate model data The large size of the difference between the estimates of the 10^7 year wind speed from the System-4 and SEAS5 seasonal forecast archives is intriguing, but also worrying: with such a large difference between the estimates from two generations of the seasonal forecast model, one may wonder whether these forecast models are mature enough to be used for the purpose of estimating extreme value statistics. It raises a difficult question: how do we decide if a weather prediction model is good enough for this purpose?

On the other hand, the problem may not be as big as it seems. Simulated near-surface wind speed is sensitive to the details of the formulation of air-sea interaction in the model, but simulated surface stress may be less sensitive. It is known that the

formulation of air-sea interaction in the two model generations is different. However, stress of System-4 is not archived, so we cannot make a direct comparison of stress. Furthermore, stress is not going to solve the problem of low regularity of the SEAS5 tail; a preliminary check showed similar irregularities in estimates of shape parameter and return values of stress as were found for wind speed in the present study.

One way to compare the models independently of their air-sea interaction formulations is to use their data as forcing of a HARMONIE mesoscale weather prediction model, for which wind profiles over sea and land have been extensively validated. In this manner, we might produce reliable wind data while benefitting from the large size of the seasonal forecast archives. However, this would require a massive computational effort. It would be more feasible to perform a limited number of HARMONIE runs, use these to learn more about the nature of bias in the System-4 and SEAS5 stress and wind speed data, and based on these insights, calibrate the seasonal forecast data.

7 Conclusions and recommendations

Conclusions

1. Accurate estimates of the wind speed with a very high return period, with root-mean square error in the order of 1-2 m/s, require a large dataset such as the ECMWF seasonal ensemble forecasts, which effectively cover 5000-6000 years. However, the value of such a dataset for this purpose depends critically on the size of bias in high wind speeds simulated by the forecast model.

2. Overall, the Generalized Weibull (GW) tail comes out as best for the purpose of estimating wind speed with very high return periods.

   (a) Estimates based on the Generalized Pareto (GP) tail have too high variance, and for some datasets, relatively large bias.

   (b) The exponential tail is positively biased; it gives the largest bias of all tails on two out of three datasets. This indicates that the estimates based on the exponential tail currently in use for flood safety assessment (Chbab, 2017) are conservative. Furthermore, confidence intervals of exponential tail fits are too narrow.

   (c) Estimates based on the GW tail are in no case much more biased than estimates based on other tails, and in two of the three datasets, bias for the GW tail is much lower. GW tail fits preserve a typical wind-surge relationship better than other tails. 1-parameter Weibull tails are less flexible, but may be preferable if estimates are very stable with varying threshold, indicating low bias.

3. The tails of wind speed of the previous System-4 and the current SEAS5 generations of the ECMWF seasonal forecasts for the position 3E, 55N in the central North Sea are very different: the tail of SEAS5 data is heavier (closer

to an exponential tail), but also less regular. This indicates that for high return periods, estimates of return values of wind speed from at least one of these models are seriously biased.

4. Ultimately, the achievable accuracy of estimates of return values is limited by the regularity of the tail. The regularity of the tail can in principle be accounted for by applying a well-established method for choosing thresholds. This is essential for obtaining the best possible estimates with realistic confidence intervals, but it is not certain that a proper method for choosing the threshold is the solution to all problems caused by irregular tails.

Recommendations

1. In order to use seasonal forecast data to determine wind speeds for very high return periods, it is necessary to find out what causes the observed differences between the tails of wind speed from the System-4 and SEAS5 datasets, and find a way to resolve these differences. Running the HARMONIE model forced by boundary conditions from System-4 and SEAS5 for a number of storms may provide much additional insight.

2. In the present study, the quantification of bias relies on the assumption of "second-order regularity" of the tails (see Appendix A). Further analysis should be carried out to check whether this assumption is justified and/or to provide an independent check of the conclusions without relying on this assumption.

3. The outcomes of the present study may already be helpful to assess and possibly improve the return values of wind speed currently in use to assess flood safety in the Netherlands (Chbab, 2017). The following steps are proposed.

   (a) Testing of the GW tail on wind measurements from different sites and over different periods in order to assess trends (see also (d) below), and comparison against estimates based on the GP and exponential tails. The 1-parameter Weibull tail may also be included in this comparison.

   (b) This should involve checking of tail regularity and a careful choice of threshold

      i. by checking the stability of estimates of the shape parameter and of return values as functions of the sample fraction used to compute the estimates;

      ii. (preferred) by a rigorous statistical analysis of fluctuations in the estimates, which helps to distinguish bias from noise.

   (c) Uncertainty assessment of wind speed for high return periods: compare the current practice (uncertainty is derived from the variance of the exponential tail fits; see Chbab (2017)) to the following alternatives:

      i. use the GW tail for return values and confidence intervals;

      ii. use the exponential tail for return values and the GW tail (or a plausible restriction of the GW model, e.g. with shape parameter $\rho \leq 1$) to estimates confidence intervals.

(d) Include interannual variability (see Section 6) in the uncertainty analysis:

i. How strong is the effect of interannual variability on uncertainty?

ii. Can uncertainty be assessed reliably in the presence of interannual variability?

References

Boucheron, S. and Thomas, M. (2015), Tail index estimation, concentration and adaptivity. *Electron. J. Statist.* **9**(2), 2751-2792.

van den Brink, H. W. and Können, G. P. (2008), The statistical distribution of meteorological outliers. *GRL* **35**, L23702, doi: 10.1029/2008GL035967.

van den Brink, H. W. en Können, G. P. (2011), Estimating 10000-year return values from short time series. *Int. J. Climatol.* **31**(1), 115–126.

van den Brink, H. W. and de Goederen (2017), Recurrence intervals for the closure of the Dutch Maeslant surge barrier. *Ocean Sci.* **13**, 691–701.

van den Brink, H.W. (2018), Extreme wind en druk in de ECMWF seizoensverwachtingen. Report, KNMI, de Bilt (in Dutch).

Caires, S. (2009), *Extreme wind statistics for the Hydraulic Boundary Conditions for the Dutch primary water defences*. Report, Deltares.

Caires. S., Groeneweg, J., van Nieuwkoop, J. (2016), Lifting of time- and space-evolving winds for the determination of extreme hydraulic conditions. *Coastal Eng.* **116**, 152–169.

Chbab, H. (2017), Basisstochasten WTI-2017 - Statistiek en statistische onzekerheid. *Report 1209433-012-HYE-0007*, Deltares, Delft (in Dutch).

Cook, N.J. (1982), Towards better estimation of extreme wind. *J. Wind Eng. Ind. Aerodyn.* **9**, 295–323.

Czörgő, M. and Révész, P. (1978), Strong Approximations of the Quantile Process. *Ann. Stat.* **6**(4), 882–894.

Dillingh, D., De Haan, L., Helmers, R., Können, G.P., Van Malde, J. (1993), De basispeilen langs de Nederlandse kust, Statistisch onderzoek, *Rapport DGW-93.023*, Rijkswaterstaat Dienst Getijdewateren (in Dutch).

Drees, H. and Kaufmann, E. (1998), Selecting the optimal sample fraction in univariate extreme value estimation. *Stochastic Process. Appl.* **75**(2), 149–172.

Drees, H. (2000), Weighted approximations of tail processes for β-mixing random variables. Ann. Appl. Prob. 10(4), 1274–1301.

Drees, H. (2003), Extreme quantile estimation for dependent data, with applications to finance. *Bernoulli* 9(4), 617–657

Ferro, C.A.T. and J. Segers (2003), Inference for clusters of extreme values. *J. R. Statist. Soc. B* **65**(2), 545–556.

Franzke, C.L.E., Osprey, S.M., Davini, P., Watkins, N.W. (2015), A Dynamical Systems Explanation of the Hurst Effect and Atmospheric Low-Frequency Variability. *Scientific Reports* **5**, article number 9068.

Furrer, E.M. and Katz, R.W. (2008), Improving the simulation of extreme precipitation events by stochastic weather generators. *Water Resources Research* **44**, W12439.

Gardes, L. and Girard, S. (2006), Comparison of Weibull tail-coefficient estimators. *REVSTAT* **4**(2), 163-188.

Hegnauer, M., Beersma, J.J., van den Boogaard, H.F.P., Buishand, T.A. and Passchier, R.H (2014), Generator of Rainfall and Discharge Extremes (GRADE) for the Rhine and Meuse basins - Final report of GRADE 2.0. *Report 1209424-004-ZWS-0018*, Deltares.

de Haan, L., Ferreira, A. (2006), *Extreme value theory - An introduction.* Springer.

Harris, R.I. (2005). Generalised Pareto methods for wind extremes. Useful tool or mathematical mirage? *J. Wind Eng. Ind. Aerodyn.* **93**, 341–360.

Justus, C.G., Hargreaves, W.R. and Yalcin, A. (1976), Nationwide Assessment of Potential Output from Wind-Powered Generators. *J. Appl. Meteor.* **15**(7), 673–678.

Kucharski, F., Molteni, F., and Bracco, A. (2006), Decadal interactions between the western tropical Pacific and the North Atlantic Oscillation. *Clim. Dyn.* **26**, 79-91

Leadbetter, M.R., Lindgren, G., RootzÃ©n, H. (1983), *Extremes and Related Properties of Random Sequences and Processes.* Springer.

Molteni, F. (2003), Atmospheric simulations using a GCM with simplified physical parametrizations. I. Model climatology and variability in multi-decadal experiments. *Clim. Dyn.* **20**, 175-191

Molteni, F. et al (2011), The new ECMWF seasonal forecast system (System 4). *Technical Memorandum 656*, ECMWF.

Rootzen, H. (1995), The tail empirical process for stationary sequences. *Preprint*, Chalmers Univ. Gothenburg.

Rootzen, H. (2008), Weak convergence of the tail empirical process for dependent sequences. *Stochastic Processes and their Applications* **119**, 468–490.

De Ronde, J.G. Van Marle, J.G.A. Roskam, A.P. Andorka Gal, J.H. (1995), Golfrandvoorwaarden langs de Nederlandse kust op relatief diep water. *Report RIKZ - 95.024*, Rijkswaterstaat RIKZ (in Dutch).

de Valk, C. (2016), Approximation of high quantiles from intermediate quantiles. *Extremes* **19**, 661–686.

de Valk, C. (2016). Approximation and estimation of very small probabilities of multivariate extreme events. *Extremes* **19**, 687–717.

de Valk, C. and Cai, J.J. (2018), A high quantile estimator based on the log-Generalised Weibull tail limit. *Econometrics and Statistics* **6**, 107–128.

De Valk, C. and Zitman, T. (1992), Determining statistics of multivariate storm extremes off the coast of the Netherlands - a pilot study. *Delft Hydraulics report H1557*, Delft.

Part II. Appendix

A Assessment of bias in quantile estimates

This section explains the method for approximating the ratio (18) , i.e.

$$\frac{\mathbb{E}(\hat{x}_{R,s} - x_R)}{\mathbb{E}(\hat{x}_{R,s} - \hat{x}_{R,f})}$$

with x_R the return value for return period R, $\hat{x}_{R,s}$ its estimate from a subsample, and $\hat{x}_{R,f}$ its estimate from the full sample.

First note that by (5), $x_R \approx Q(p_R)$ with p_R given by (6). Therefore, the estimates $\hat{x}_{R,s}$ and $\hat{x}_{R,f}$ are really estimates of the quantile $Q(p_R)$ for a fraction of time p_R. Denote the estimate of $Q(p_R)$ from a subsample by $\hat{Q}_2(p_R)$, and the estimate from the full sample by $\hat{Q}_1(p_R)$. Then

$$\frac{\mathbb{E}(\hat{x}_{R,s} - x_R)}{\mathbb{E}(\hat{x}_{R,s} - \hat{x}_{R,f})} \approx \frac{\mathbb{E}(\hat{Q}_2(p_R) - Q(p_R))}{\mathbb{E}(\hat{Q}_2(p_R) - \hat{Q}_1(p_R))}. \tag{20}$$

The quantile estimates $\hat{Q}_2(p_R)$, and $\hat{Q}_1(p_R)$ for the different types of tails are just the quantile approximations (12), (13), (14), (15) and (16) in Chapter 1, with all their parameters replaced by estimates. In fact, the expectations $\mathbb{E}\hat{Q}_2(p_R)$ and $\mathbb{E}\hat{Q}_1(p_R)$ of these quantile estimates look like the quantile approximations in Chapter 1, but contain additional terms representing

1. the deviation between Q and its approximation (a second-order approximation)

2. if a shape parameter is estimated: a term representing the effect of shape estimation on the expectation of the quantile estimate.

As in (12), (13), (14), (15) and (16), $\mathbb{E}\hat{Q}_2(p_R)$ and $\mathbb{E}\hat{Q}_1(p_R)$ are functions of p_R and of a threshold probability larger than p_R. This threshold probability is larger for the estimate $\hat{Q}_2(p_R)$ from a subsample than for the estimate $\hat{Q}_1(p_R)$ from the full sample.

In the following sections, approximations are given of the ratio on the right-hand side of (20), which are determined by the type of tail considered, and by the threshold probabilities for the estimates $\hat{Q}_1(p_R)$ and $\hat{Q}_2(p_R)$. These approximations take the form of limits obtained by letting p_R tend to zero. For simplicity, the symbol p_R will be replaced by the generic symbol p.

A.1 GW tail

The second-order model for the log-GW tail limit (10), (15) as proposed in de Valk & Cai (2018) (eq. (6) and (9)-(12)) can be translated straightforwardly to the GW case. Using the terminology of Section 2, define

$$q(y) := Q(e^{-y})$$

and for every real number α,

$$h_\alpha(\lambda) := \frac{1}{\alpha}(\lambda^\alpha - 1), \qquad \lambda > 0,$$

to be interpreted as $\log \lambda$ if $\alpha = 0$.

Assume that (a) q is absolutely continuous with derivative q' and (b) q' is regulary varying (de Haan & Ferreira, 2006): for some positive function φ,

$$\lim_{y\uparrow\infty} \frac{q'(y\lambda)}{q'(y)} = \varphi(\lambda), \qquad \lambda > 0.$$

Define

$$f(p) := (-\log p)q'(-\log p);$$

then the assumption above implies that for some real number ρ,

$$\lim_{p\downarrow 0} \frac{f(p^\lambda)}{f(p)} = \lambda \lim_{y\uparrow\infty} \frac{q'(y\lambda)}{q'(y)} = \lambda^\rho, \qquad \lambda > 0, \tag{21}$$

and by integration, we obtain the GW tail limit (14).

Now we strengthen (21) with the following second-order regularity assumption (cf. (9) in de Valk & Cai (2018) for the log-GW tail limit): for some non-constant function Ω and some function η satisfying that $\lim_{p\downarrow 0} \eta(p) = 0$,

$$\frac{\lambda^{-\rho}\frac{f(p^\lambda)}{f(p)} - 1}{\eta(p)} = (1 + o(1))\Omega(\lambda) \tag{22}$$

locally uniformly in $\lambda > 0$. This implies that (de Valk & Cai, 2018, Section 2)

$$\lim_{p\downarrow 0} \frac{\eta(p^\lambda)}{\eta(p)} = \lambda^\delta(1 + o(1)), \qquad \lambda > 0 \tag{23}$$

and furthermore, $\Omega(\lambda) = h_\delta(\lambda)$ for some $\delta \le 0$, and

$$\frac{Q(p^\lambda) - Q(p)}{f(p)} = h_\rho(\lambda) + (1 + o(1))\eta(p)\Psi_{\rho,\delta}(\lambda) \tag{24}$$

locally uniformly in $\lambda > 0$, with

$$\Psi_{\rho,\delta}(\lambda) := \frac{1}{\delta}(h_{\rho+\delta}(\lambda) - h_\rho(\lambda))$$

(for $\delta = 0$ to be interpreted as its limit as $\delta \uparrow 0$).

Let $\hat{Q}_i(p)$ be the estimator of the quantile $Q(p)$ with threshold at an order statistic exceeded by a fraction p^{1/λ_i} of the sample (see Appendix B), so the threshold is approximately $Q(p^{1/\lambda_i})$. By analogy to the log-GW case, we obtain the following approximations of its expectation $\mathbb{E}\hat{Q}_i(p)$ from Theorem 3 in de Valk & Cai (2018):

$$\mathbb{E}\hat{Q}_i(p) = Q(p^{1/\lambda_i}) + f(p^{1/\lambda_i})(h_\rho(\lambda_i) + (1+o(1))\Psi_{\rho,0}(\lambda_i)\eta(p^{1/\lambda_i})) \qquad p \in (0,1), i \in \{1,2\}.$$

From (24):

$$Q(p) = Q(p^{1/\lambda_i}) + f(p^{1/\lambda_i})(h_\rho(\lambda_i) + (1+o(1))\Psi_{\rho,\delta}(\lambda_i)\eta(p^{1/\lambda_i})),$$

and combined with the previous,

$$\mathbb{E}\hat{Q}_i(p) - Q(p) = f(p^{1/\lambda_i})\eta(p^{1/\lambda_i})(\Psi_{\rho,0}(\lambda_i) - \Psi_{\rho,\delta}(\lambda_i) + o(1)). \tag{25}$$

This expression shows that the error term

$$f(p^{1/\lambda_i})\Psi_{\rho,\delta}(\lambda_i)\eta(p^{1/\lambda_i})$$

in the second-order approximation of $Q(p)$ is partially cancelled by the term

$$f(p^{1/\lambda_i})\Psi_{\rho,0}(\lambda_i)\eta(p^{1/\lambda_i}),$$

which represents bias in $\mathbb{E}\hat{Q}_i(p)$ due to bias in the shape parameter estimator.

Suppose that $\delta < 0$. By (25), (21) and (23), we obtain

$$\frac{\mathbb{E}\hat{Q}_2(p) - Q(p)}{\mathbb{E}\hat{Q}_1(p) - Q(p)} = (1+o(1))(\lambda_1/\lambda_2)^{\rho+\delta}\frac{\Psi_{\rho,0}(\lambda_2) - \Psi_{\rho,\delta}(\lambda_2)}{\Psi_{\rho,0}(\lambda_1) - \Psi_{\rho,\delta}(\lambda_1)}$$

and therefore, the ratio (20) can be approximated for small p by the limit

$$\lim_{p\downarrow 0}\frac{\mathbb{E}\hat{Q}_1(p) - Q(p)}{\mathbb{E}\hat{Q}_2(p) - \mathbb{E}\hat{Q}_1(p)} = \left((\lambda_1/\lambda_2)^{\rho+\delta}\frac{\Psi_{\rho,0}(\lambda_2) - \Psi_{\rho,\delta}(\lambda_2)}{\Psi_{\rho,0}(\lambda_1) - \Psi_{\rho,\delta}(\lambda_1)} - 1\right)^{-1}. \tag{26}$$

The ratio (26) increases with increasing $\delta < 0$. It is undetermined if $\delta = 0$, because the numerator and denominator both vanish in that case. Therefore, we consider the limit of the right-hand side of (26) as $\delta \uparrow 0$ as an upper bound, which should be good enough for a crude assessment of worst-case bias.

A.2 log-GW tail

For the log-GW tail limit (10), the second-order model is

$$\frac{\lambda^{-\theta\frac{g(p^\lambda)}{g(p)}} - 1}{\eta(p)} = (1+o(1))h_\delta(\lambda)$$

and

$$\frac{\log Q(p^\lambda) - \log Q(p)}{g(p)} = h_\theta(\lambda) + (1+o(1))\eta(p)\Psi_{\theta,\delta}(\lambda)$$

for some $\delta \le 0$ (compare (22) and below).

In this case, assuming again that $\delta < 0$, we obtain in a similar manner,

$$\lim_{p\downarrow 0}\frac{\mathbb{E}\log\hat{Q}_1(p) - \log Q(p)}{\mathbb{E}\log\hat{Q}_2(p) - \mathbb{E}\log\hat{Q}_1(p)} = \left((\lambda_1/\lambda_2)^{\theta+\delta}\frac{\Psi_{\theta,0}(\lambda_2) - \Psi_{\theta,\delta}(\lambda_2)}{\Psi_{\theta,0}(\lambda_1) - \Psi_{\theta,\delta}(\lambda_1)} - 1\right)^{-1}. \tag{27}$$

with θ the log-GW shape parameter as in (10) and δ the second-order shape parameter for the log-GW tail limit.

Furthermore, assuming that g in (10) is of bounded increase (which is the case if a GW tail limit applies, for example), then (26) with ρ replaced by $\theta \le 0$ holds.

A.3 Exponential tail

The exponential tail is a special case of the GW tail (a GW tail with $\rho = 1$) as well as of the GP tail (a GP tail with $\gamma = 0$). Based on the same second-order model as in Section A.1 and assuming that $\delta < 0$, we obtain

$$\lim_{p\downarrow 0} \frac{\mathbb{E}\hat{Q}_1(p) - Q(p)}{\mathbb{E}\hat{Q}_2(p) - \mathbb{E}\hat{Q}_1(p)} = \left((\lambda_1/\lambda_2)^{1+\delta} \frac{\Psi_{1,\delta}(\lambda_2)}{\Psi_{1,\delta}(\lambda_1)} - 1\right)^{-1}. \tag{28}$$

The difference with (26) in the case $\rho = 1$ is due to the fact that ρ is not estimated when fitting the exponential tail, so there is no partial cancellation of bias.

A.4 1-parameter Weibull tail

The 1-parameter Weibull tail is a special case of the log-GW tail (with $\theta = 0$). Based on a second-order model for the log-GW tail (similar to the second-order model for the GW tail in Section A.1) and assuming that $\delta < 0$, we obtain (as g in (10) is of bounded increase; see Subsection A.2):

$$\lim_{p\downarrow 0} \frac{\mathbb{E}\hat{Q}_1(p) - Q(p)}{\mathbb{E}\hat{Q}_2(p) - \mathbb{E}\hat{Q}_1(p)} = \lim_{p\downarrow 0} \frac{\mathbb{E}\log\hat{Q}_1(p) - \log Q(p)}{\mathbb{E}\log\hat{Q}_2(p) - \mathbb{E}\log\hat{Q}_1(p)} = \left((\lambda_1/\lambda_2)^{\delta} \frac{\Psi_{0,\delta}(\lambda_2)}{\Psi_{0,\delta}(\lambda_1)} - 1\right)^{-1}. \tag{29}$$

As in the case of the exponential tail, there is no partial cancellation of bias because the shape is not estimated.

A.5 GP tail

Finally, a similar analysis can be made for the GP tail under the assumption that the GP tail limit applies. From the second-order model for the GP tail limit (7) as in de Haan & Ferreira (2006) (e.g. Section 2.3):

$$Q(p) = Q(p\xi_i) + a(p\xi_i)(h_\gamma(\xi_i) + (1 + o(1))\Psi_{\gamma,\delta}(\xi_i)\eta(p\xi_i)), \quad \xi_i > 0.$$

with a and γ the scale and shape parameter of the GP tail limit, respectively.

A major difference with the previous models is that $Q(p)$ is extrapolated from $Q(p\xi_i)$ instead of from $Q(p^{1/\lambda_i})$. Setting $p\xi_i = p^{1/\lambda_i}$ (i.e., extrapolating from the same thresholds), we obtain $\xi_i = p^{1/\lambda_i - 1}$. Even for moderate $\lambda_i > 1$, ξ_i can be a very large number if p is very small, which is the case we are interested in.

From de Haan & Ferreira (2006) (in particular: eq. (3.4.6), (3.4.7) and the proof of Theorem 4.3.1), assuming that the second-order parameter of the GP tail limit δ is negative, we can obtain the following error approximations for the estimator of a high quantile based on the MLE estimator of the scale and shape parameters of the GP tail limit:

$$\begin{aligned} &\mathbb{E}\hat{Q}_i(p) - Q(p) \\ &= a(p\xi_i)\eta(p\xi_i)\left(\frac{\Psi_{\gamma,0}(\xi_i)(1+\gamma) - h_\gamma(\xi_i)\delta}{(1-\delta)(1+\gamma-\delta)}(1+o(1)) - \Psi_{\gamma,\delta}(\xi_i)(1+o(1))\right). \end{aligned} \tag{30}$$

Using the regular variation of a and η, this gives

$$\frac{\mathbb{E}\hat{Q}_2(p) - Q(p)}{\mathbb{E}\hat{Q}_1(p) - Q(p)}$$
$$= (1 + o(1))(\xi_1/\xi_2)^{\gamma+\delta} \frac{\Psi_{\gamma,0}(\xi_2)(1+\gamma) - h_\gamma(\xi_2)\delta - \Psi_{\gamma,\delta}(\xi_2)(1-\delta)(1+\gamma-\delta) + o(1)}{\Psi_{\gamma,0}(\xi_1)(1+\gamma) - h_\gamma(\xi_1)\delta - \Psi_{\gamma,\delta}(\xi_1)(1-\delta)(1+\gamma-\delta) + o(1)}$$

and therefore,

$$\lim_{p\downarrow 0} \frac{\mathbb{E}\hat{Q}_1(p) - Q(p)}{\mathbb{E}\hat{Q}_2(p) - \mathbb{E}\hat{Q}_1(p)} \tag{31}$$
$$= \left((\xi_1/\xi_2)^{\gamma+\delta} \frac{\Psi_{\gamma,0}(\xi_2)(1+\gamma) - h_\gamma(\xi_2)\delta - \Psi_{\gamma,\delta}(\xi_2)(1-\delta)(1+\gamma-\delta)}{\Psi_{\gamma,0}(\xi_1)(1+\gamma) - h_\gamma(\xi_1)\delta - \Psi_{\gamma,\delta}(\xi_1)(1-\delta)(1+\gamma-\delta)} - 1 \right)^{-1}.$$

In our application to the estimation of the 10^7-year wind speed, ξ_1 and ξ_2 are very large numbers. As long as $\delta < 0$, this would not invalidate the analysis above. Again, for this expression, we only consider the limit upon $\delta \uparrow 0$, as a plausible worst case.

B Estimators for the log-GW and GW tails

Starting point for estimation of the log-GW and GW tails was the estimator for the log-GW tail (10) proposed in de Valk & Cai (2018), because the large-sample behaviour of this estimator has been thoroughly analyzed. The estimator is based on a second-order tail regularity assumption as in Section A.1. It uses different thresholds for estimation of the scale and shape parameter, which is necessary to obtain consistency without assumptions on rate of convergence to the tail limit. The threshold for the shape parameter is lower (and is therefore exceeded by a larger number of order statistics) than the threshold for the scale parameter.

In the present study, we used a variant of this estimator. In the notation of de Valk & Cai (2018), the estimator $\hat{\theta}_{k_n,n}$ of the shape parameter θ in eq. (19) of de Valk & Cai (2018) has been modified by replacing $\hat{\gamma}^H_{i,n}$ by

$$\left(\vartheta_{i+1,n} \frac{1}{i} \sum_{j=1}^{i} h_{\hat{\theta}_{k_n,n}}(\vartheta_{j,n}/\vartheta_{i+1,n}) \right)^{-1} \hat{\gamma}^H_{i,n}.$$

This modification is a result of replacing the first approximation of $g(\vartheta_{i+1,n})$ by the second approximation in (17) of de Valk & Cai (2018). It leads to an equation to be solved for $\hat{\theta}_{k_n,n}$. It can be shown that this estimator has the same asymptotic statistics as the original estimator of de Valk & Cai (2018).

Given the threshold for the scale parameter, the parameter $\sigma > 0$ of the estimator controls the threshold for the shape parameter; see eq. (30) in de Valk & Cai (2018). A lower value of σ gives a lower threshold for the shape, exceeded by a larger number of order statistics. This reduces the variance of the shape estimator and the quantile estimator, but it may increase bias. In the examples in de Valk & Cai (2018), $\sigma = 1$ was used. The present study used $\sigma = 2$: this gives a somewhat higher variance of quantile estimates, but it reduces bias when convergence to a tail limit is slow, as appears to be the case of SEAS5 data (see Section 4.4).

For the GW tail, the same estimator can be used after replacing the data by their exponents. Because the estimator for the log-GW tail above is an operator on logarithms of the data, this amount to skipping a logarithmic transformation.

In addition, a modified maximum likelihood estimator (MLE) for the GW and log-GW tails was implemented and tested. The modification involves estimating the scale and shape parameter at different thresholds just as in de Valk & Cai (2018): the MLE's are computed at both thresholds, and then the scale estimator at the higher threshold is retained, while the shape parameter estimator at the lower threshold is retained. Experiments with this modified MLE for the GW and log-GW tails showed that estimates are almost identical to those of the modified de Valk & Cai (2018) estimator described above. Therefore, it is a good guess that the asymptotics of this modified MLE are the same as for the modified de Valk & Cai (2018) estimator. In the present study, only the latter was used, because its computation is much faster and its asymptotics are known. The modified MLE has the advantage that it readily extends to models with the tail depending on covariates such as time, location and wind direction.

Royal Netherlands Meteorological Institute

PO Box 201 | NL-3730 AE De Bilt
Netherlands | www.knmi.nl