International Journal of Disaster Risk Science
https://doi.org/10.1007/s13753-026-00764-4

www.ijdrs.com
www.springer.com/13753

PERSPECTIVE

On the Reliability of Estimated Return Periods for Climate Extremes

Elisa Ragno[1] · Amir AghaKouchak[2]

Accepted: 19 August 2026

Abstract

This study reflects on the probability of observing an extreme event of interest within a finite dataset, whether derived from observations or model simulations, to inform risk assessment or climate adaptation efforts. To do so, we adopt the concept of engineering reliability, which is defined as the probability that a system remains in a satisfactory state, to assess the reliability of extreme events inferred from a dataset, whether this is from observations or model simulations. This assessment links the number of available observations or simulations to the low frequency of the event, providing a quantitative measure of confidence in our ability to observe or simulate such events over a given time horizon. This approach offers a fresh perspective on the interpretation of an extreme event, where the rarity of an event is considered not only in terms of its frequency but also relative to the length of the dataset used. Our reflections aim to guide preparedness for future extremes and highlight the scientific challenges inherent in their prediction and projection. We emphasize that while large ensembles are essential to overcome the limitations of historical observations, they should be used with caution to avoid overconfidence arising from underlying modeling assumptions. Finally, we stress that statistical extrapolation, whether it is parametric or non-parametric, is unavoidable, as the link between event frequency and the definition of extremes cannot be eliminated.

Keywords Dataset length · Extremes · Large ensembles · Reliability · Return period

1 Introduction

Urban areas remain vulnerable to climate extremes due to a combination of more frequent and severe hazards, for example, flooding, storms, droughts (IPCC 2023), and greater exposure, that is, population growth and infrastructure expansion (McPhillips et al. 2018; Raju et al. 2022).

The conflict between climate and human settlements is not new. Already in 1976, after the UN Conference on Human Settlements (Habitat), the World Meteorological Organization (WMO) released a report describing the role of houses, for example, in protecting people against adverse weather and the use of weather statistics to determine design values enhancing security and minimizing human disasters (Landsberg 1976).

✉ Elisa Ragno
e.ragno@tudelft.nl

1 Faculty of Civil Engineering and Geosciences, Delft University of Technology, 2628 CN Delft, Netherlands

2 Department of Civil and Environmental Engineering, University of California, Irvine, CA 92697, USA

The traditional use of weather statistics hinges on the natural variability of the climate (Salas and Obeysekera 2014). For this, the occurrence of an extreme event, such as a flood event, is treated as a random variable (Beven 2013) described by a probability distribution function fitted to available observations of the phenomenon of interest. In engineering design and risk assessment, buildings and critical infrastructure are designed to withstand the expected magnitude of one or more extreme weather events, that is, the expected magnitude of an event that, if it were to occur, might cause the failure of a structure. The magnitude of such an event is often extrapolated from the tail of a probability distribution function, that is, it has a low probability of being exceeded, and often it has not been observed in the past. The inferred event is often referred to as a T-year event, expected (but not necessarily) to occur, on average, and over a long period, once every T-years. For example, a 100-year event, or an event with a return period of 100 years, is an event that is expected to occur on average once every 100 years or has a 0.01 annual probability of being exceeded (see Volpi et al. (2015) and references therein).

Recent studies have suggested the use of large ensembles to not only investigate the properties of weather

Published online: 03 September 2026

extremes in current and future climate but also to obtain the magnitude of extreme events for design and risk assessment purposes (Leander et al. 2005; Gitau et al. 2018; van der Wiel et al. 2019; Kelder et al. 2020; Raynaud et al. 2020; Bevacqua et al. 2023). Large ensembles consist of multiple simulations under the same climate forcing, generating diverse physically plausible realizations that allow assessment of stochastic internal climate variability and provide data beyond climate model simulations like CIMP6 (Maher et al. 2021; Lehner and Deser 2023). Hence, large ensembles overcome the limited availability of observations while providing physically plausible weather events (Thompson et al. 2017). For example, Fischer et al. (2023) showed that an ensemble boosting approach can generate physically plausible storylines of a heatwave hotter than the one observed in the Pacific Northwest in late June 2021, which was not predictable based on observations only. This approach is still difficult to use in practice, for example, it is challenging to assign a frequency to extreme events of this magnitude. However, the use of conditional probability seems promising when quantifying the return period of such extreme events (Bloin-Wibe et al. 2025). Generally speaking, there is no accepted blueprint on how large ensembles can be used to evaluate current and future climate (Vavrus et al. 2015), even though effort has been made in identifying the most suitable ensemble size for a representative ensemble, which is often purpose-specific (Milinski et al. 2020; Tebaldi et al. 2021).

In this study, we want to reflect on the probability of observing an extreme event of interest within the available dataset, be it derived from observations or model simulations, to help guide climate adaptation actions. Here, we use the concept of engineering reliability, that is, the probability of the system to remain in a satisfactory state (Hashimoto et al. 1982), to assess the reliability of extreme events inferred from a dataset, regardless of whether this inference is done from a parametric or non-parametric distribution, by linking the number of observations used and the (low) frequency of the event of interest. In doing so, we quantify a level of confidence regarding the ability to observe or simulate the event of interest (for example, an event with low frequency) over the observation or simulation horizon (Read and Vogel 2015), leading to a fresh perspective on the meaning of “extreme event.” In this context, the extreme character of the event is determined not only by its frequency but also in relation to the length of observations/simulations used.

In the remainder of the article, we introduce the concept of reliability. We then use it to determine the reliability of the event of interest obtained from a dataset (observed or simulated). We then reflect on how this different perspective can prevent overconfidence when using large model ensembles.

2 Reliability of Extremes via Poisson Distribution

In engineering, reliability is defined as the probability that a system will remain in a satisfactory state over its lifetime (Hashimoto et al. 1982). The reliability, R_N, of a system, for example, infrastructure or water project, over its lifespan (N in years), assuming a T-year design event, is written as

$$R_N = \left(1 - \frac{1}{T}\right)^N \tag{1}$$

where $1/T$ is the yearly probability that an event with the same or greater magnitude can occur. Equation 1 is valid under the assumption that the loads in each year are independent and identically distributed, see Vogel and Castellarin (2017) and references therein. In other words, R_N quantifies the confidence that the system is safe over its lifespan, which is the probability that an event equal or greater than the design event will not occur, under the assumption that system failure is determined solely by the hazard. However, this equation does not indicate whether the infrastructure or system would physically fail if an extreme event exceeds the design threshold.

Equation 1 can also be interpreted as the probability that the T-year event will not be observed over a period of N-years. If we take the complement, R_O, of the engineering reliability, we obtain the probability of observing the event of interest at least once.

$$R_O = 1 - \left(1 - \frac{1}{T}\right)^N \tag{2}$$

In other words, R_O tells us how confident we are that the T -year event is included in our dataset of observations/simulations. To relate the T-year (extreme) event of interest to the dataset size N, we define the event as a function of N where α is the ratio between the return period T in years and dataset size N in years ($T = \alpha N$).

$$R_O = 1 - \left(1 - \frac{1}{\alpha N}\right)^N \tag{3}$$

By taking the limit of Eq. 3 for $N \to \infty$, we can express the reliability of the estimate via a Poisson distribution (Yen 1970).

$$R_O \sim \lim_{N\to\infty} 1 - \left(1 - \frac{1}{\alpha N}\right)^N = 1 - e^{-1/\alpha} \tag{4}$$

The main assumption behind the metric R_O is that the probability of having observed the event of interest in the available dataset is well described by the Poisson distribution, Eq. 4. The Poisson distribution is the limiting case of a Binomial distribution $Bin(n, p)$ for $N \to \infty$ where $n = N$ is the number of years of observations/simulations and

$p = 1/T$ is the frequency of the event of interest (Dekking et al. 2005). The advantage of assuming the Poisson distribution resides in R_O being independent of the underlying distribution function used to model the phenomenon of interest. However, this approximation does not hold for small dataset sizes ($N < 30$) and large return periods (for example, $T >> N$).

Following R_O, if we are interested in a 30-year event ($T = 30$) and we have 100 years of data ($N = 100$), we can confidently say that the event of interest was included in the observations/simulations available (Fig. 1 green panel). On the other hand, if we would like to infer the 500-year event ($T = 500$) from only 50 years of observation or simulations ($N = 50$), we will have very little confidence of having observed/simulated this event in the dataset (Fig. 1 red panel). In those cases where the event of interest is very similar to the number of observations, it is more difficult to say whether the event has been observed or generated (Fig. 1 yellow panel). More specifically, when $N = T$ and so $\alpha = T/N = 1$, the probability of observing or simulating the event of interest is always equal to 0.63 (Yen 1970). This implies that the probability of observing a 30-year event ($T = 30$) over a 30-year period ($N = 30$) is the same as the probability of observing a 1000-year ($T = 1000$) event over a 1000-year period ($N = 1000$). Moreover, this clearly shows that, from the perspective of the reliability of extremes, the extreme character of the event of interest should be intended in relation to the observations or simulations used to derive it and not merely in its absolute terms.

3 Reflecting on the Reliability of Extremes

From a risk management perspective, past experiences strongly influence future actions and policy plans. For example, the 1953 storm surge flood in the Netherlands serves as an example since it shaped current flood management practices in the country. However, preparedness against future extremes means going beyond historical events to prevent future human disasters and high-impact events. It is important to accept the fact that the past might not be a good guide to the future, especially with the evidence of more frequent and severe extremes. Moreover, unexpected events are intrinsic to nonlinear and dynamic systems, and/or can be caused by unusual or unexpected boundary forcing (Beven 2013). The length of recorded observations is on average around 30 to 50 years (with some exceptions). This implies that the inference, for example, the 100-year event, is often outside the range of observations and heavily relies on the selected statistical model. Following the concept of the reliability of extremes proposed here, the probability of observing a 100-year event in 50 years of observations is about 0.40 and reduces to 0.26 in the case of 30 years of observations. In this context, the advances in the use of global climate models to simulate physically plausible but previously unobserved events are particularly important, as shown by Fischer et al. (2023) and Bloin-Wibe et al. (2025) in the case of the Pacific Northwest heatwave in late June 2021.

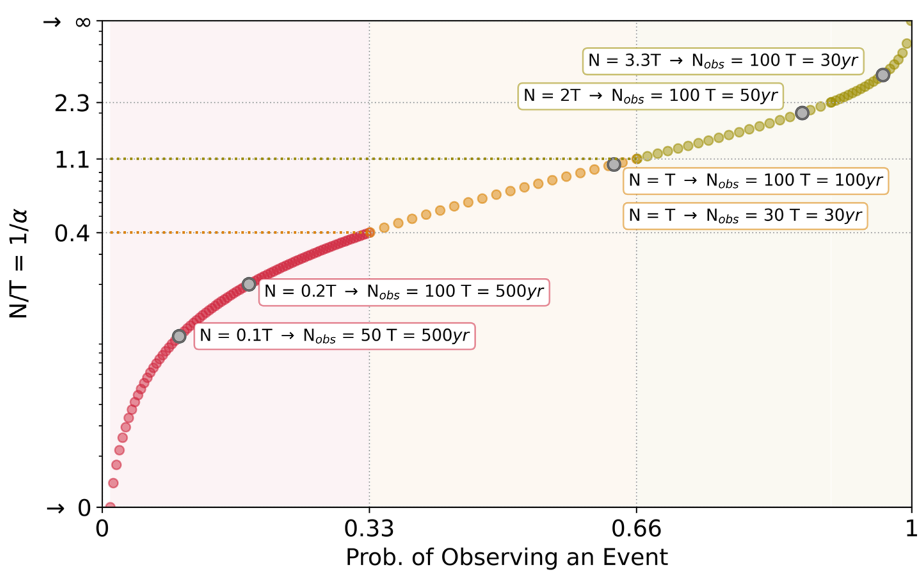

Fig. 1 Probability of having observed the event of interest via the Poisson distribution as a function of the dataset size (observed or simulated). More specifically, the figure maps the probability of observing the event with a return period T (x-axis) to the ratio between the number of observations/simulations available (N) and the return period (T) (y-axis). The relationship between T and N assumed is T = αN to explicitly link the severity of an event with the length of the data used. The background colors range from red, meaning low probability of observing the event of interest in the sample available, to green, high probability of observing the event

It is worth noticing that capturing internal variability in climate models is more difficult compared to the models' response to external forcings (Milinski et al. 2020; Tebaldi et al. 2021). Moreover, the computational demand required to produce a large ensemble, as well as the validation of its representativeness, make this approach not always viable.

Large ensembles do solve the issue of limited data availability, assuming that these ensembles are representative of the climate and its evolution. At the same time, though, ensemble members are generated from climate models that have been validated against observations. That is why we trust them. Model validation consists of checking that the statistical properties of the simulated climate match those of the observations, as seen, for example, in the cases of Thompson et al. (2017) and van der Wiel et al. (2019). The reduction in uncertainty therefore arises from the larger number of events available in the model simulations compared to the observational record. This might lead to overconfidence in the results obtained and does not necessarily provide better estimates. Indeed, the concept of reliability of extremes for large ensembles becomes even more interesting because the probability of observing/simulating the event with a return period equal to the length of the dataset is always 0.63 and does not increase with the dataset size. For this, extremes should be considered in relation to the observations or simulations used to derive them and not in their absolute terms.

Moreover, at first glance, the use of large ensembles suggests the possibility of avoiding statistical extrapolation. However, this is not the case because the severity of an event is expressed in terms of its return period, that is, the frequency of being exceeded. A nonparametric option is to make use of plotting positions, which, however, implies empirical interpolation of the frequency of the events, in this case, simulated. However, depending on which approach is selected, a different return period (or probability of occurrence) is assigned to it, implying a sort of empirical extrapolation (Makkonen 2006). An alternative option is to employ order statistics. However, in such a case it has been shown that the return period $T_{r,N}$ associated to the r^{th} order statistics, where $r = 1$ represents the largest observations in a dataset of length N is equal to $N/(r-1)$ and it is only defined for $r \geq 2$ (Vogel and Castellarin 2017). This implies that the return period of the highest event cannot be obtained since it reaches infinity.

The considerations made so far assume that the events investigated are stationary. In the presence of nonstationarity, Eq. 4 does not hold any longer because each year the probability associated with an event with fixed magnitude changes. Generally speaking, inferences are challenging because the statistical properties of the event of interest are not constant. Some authors suggested ways to determine return periods under the nonstationary assumptions, for example, Salas and Obeysekera (2014); while others, such as Read and Vogel (2015), suggested avoiding the use of return periods in favor of the concept of reliability. Read and Vogel (2015) proposed to fix a desired level of reliability within the horizon of the design project and determine numerically the design values. Following this idea, in the context of the reliability of extremes, R_O can be obtained similarly, allowing the exceedance probability of an event of fixed magnitude to vary. Hence, the inferred event will have a reliability level attached to it to inform the decision-making process.

4 Conclusion

The reflection made here is an attempt to guide preparedness for future extremes and to shed light on scientific challenges associated with the inference of future extreme events. We argue that information on the reliability of extreme events, which are often perceived as certain estimates, can help guide towards the use of large ensembles, or better, towards a critical use of the available resources, whether they come from observations or model simulations. Large ensembles are extremely valuable to compensate for the limited availability of observations. At the same time, they should be used with caution, as they can induce a false sense of confidence in our estimates, which may stem primarily from underlying modeling assumptions and biases. Moreover, we argue that (parametric or nonparametric) statistical extrapolation is unavoidable, as the connection between event frequency cannot be eliminated from the definition of extreme event.

Acknowledgements The authors would like to thank the editors and the reviewers for their valuable time.

References

Bevacqua, E., L. Suarez-Gutierrez, A. Jézéquel, F. Lehner, M. Vrac, P. Yiou, and J. Zscheischler. 2023. Advancing research on compound weather and climate events via large ensemble model simulations. *Nature Communications* 14(1): Article 2145.

Beven, K. 2013. So how much of your error is epistemic? Lessons from Japan and Italy. *Hydrological Processes* 27(11): 1677–1680.

Bloin-Wibe, L., R. Noyelle, V. Humphrey, U. Beyerle, R. Knutti, and E. Fischer. 2025. Estimating return periods for extreme events in climate models through ensemble boosting. *Weather and Climate Dynamics* 6(4): 1147–1177.

Dekking, F.M., C. Kraaikamp, H.P. Lopuhaä, and L.E. Meester. 2005. *A modern introduction to probability and statistics: Understanding why and how*. London: Springer.

Fischer, E.M., U. Beyerle, L. Bloin-Wibe, C. Gessner, V. Humphrey, F. Lehner, A.G. Pendergrass, S. Sippel, et al. 2023. Storylines for unprecedented heatwaves based on ensemble boosting. *Nature Communications* 14(1): Article 4643.

Gitau, M.W., S. Mehan, and T. Guo. 2018. Weather generator effectiveness in capturing climate extremes. *Environmental Processes* 5: 153–165.

Hashimoto, T., J.R. Stedinger, and D.P. Loucks. 1982. Reliability, resiliency, and vulnerability criteria for water resource system performance evaluation. *Water Resources Research* 18(1): 14–20.

IPCC (Intergovernmental Panel on Climate Change). 2023. Summary for policymakers. In *Climate change 2021: The physical science basis. Contribution of Working Group I to the Sixth Assessment Report of the Intergovernmental Panel on Climate Change*, ed. V. Masson-Delmotte, P. Zhai, A. Pirani, S.L. Connors, C. Péan, S. Berger, N. Caud, Y. Chen, et al., 3–32. Cambridge, UK: Cambridge University Press.

Kelder, T., M. Müller, L.J. Slater, T.I. Marjoribanks, R.L. Wilby, C. Prudhomme, P. Bohlinger, L. Ferranti, and T. Nipen. 2020. Using UNSEEN trends to detect decadal changes in 100-year precipitation extremes. *npj Climate and Atmospheric Science* 3(1): Article 47.

Landsberg, H.E. 1976. Special Environmental Report No. 7: Weather, climate and human settlements. Geneva: World Meteorological Organization.

Leander, R., A. Buishand, P. Aalders, and M.D. Wit. 2005. Estimation of extreme floods of the River Meuse using a stochastic weather generator and a rainfall-runoff model. *Hydrological Sciences Journal* 50(6): 1089–1103.

Lehner, F., and C. Deser. 2023. Origin, importance, and predictive limits of internal climate variability. *Environmental Research: Climate* 2(2): Article 023001.

Maher, N., S. Milinski, and R. Ludwig. 2021. Large ensemble climate model simulations: Introduction, overview, and future prospects for utilising multiple types of large ensemble. *Earth System Dynamics* 12(2): 401–418.

Makkonen, L. 2006. Plotting positions in extreme value analysis. *Journal of Applied Meteorology and Climatology* 45(2): 334–340.

McPhillips, L.E., H. Chang, M.V. Chester, Y. Depietri, E. Friedman, N.B. Grimm, J.S. Kominoski, and T. McPhearson et al. 2018. Defining extreme events: A cross-disciplinary review. *Earth's Future* 6(3): 441–455.

Milinski, S., N. Maher, and D. Olonscheck. 2020. How large does a large ensemble need to be?. *Earth System Dynamics* 11(4): 885–901.

Raju, E., E. Boyd, and F. Otto. 2022. Stop blaming the climate for disasters. *Communications Earth & Environment* 3(1): Article 1.

Raynaud, D., B. Hingray, G. Evin, A.C. Favre, and J. Chardon. 2020. Assessment of meteorological extremes using a synoptic weather generator and a downscaling model based on analogues. *Hydrology and Earth System Sciences* 24(9): 4339–4352.

Read, L.K., and R.M. Vogel. 2015. Reliability, return periods, and risk under nonstationarity. *Water Resources Research* 51(8): 6381–6398.

Salas, J.D., and J. Obeysekera. 2014. Revisiting the concepts of return period and risk for nonstationary hydrologic extreme events. *Journal of Hydrologic Engineering* 19: 554–568.

Tebaldi, C., K. Dorheim, M. Wehner, and R. Leung. 2021. Extreme metrics from large ensembles: Investigating the effects of ensemble size on their estimates. *Earth System Dynamics* 12(4): 1427–1501.

Thompson, V., N.J. Dunstone, A.A. Scaife, D.M. Smith, J.M. Slingo, S. Brown, and S.E. Belcher. 2017. High risk of unprecedented UK rainfall in the current climate. *Nature Communications* 8(1): Article 107.

van der Wiel, K., N. Wanders, F.M. Selten, and M.F.P. Bierkens. 2019. Added value of large ensemble simulations for assessing extreme river discharge in a 2 C warmer world. *Geophysical Research Letters* 46(4): 2093–2102.

Vavrus, S.J., M. Notaro, and D.J. Lorenz. 2015. Interpreting climate model projections of extreme weather events. *Weather and Climate Extremes* 10: 10–28.

Vogel, R.M., and A. Castellarin. 2017. Risk, reliability, and return periods and hydrologic design. In *Handbook of applied hydrology*, ed. V.P. Singh, Chapter 78. New York: McGraw-Hill.

Volpi, E., A. Fiori, S. Grimaldi, F. Lombardo, and D. Koutsoyiannis. 2015. One hundred years of return period: Strengths and limitations. *Water Resources Research* 51(10): 8570–8585.

Yen, B.C. 1970. Risks in hydrologic design of engineering projects. *Journal of the Hydraulics Division* 96(4): 959–966.