Estimating return periods for extreme events in climate models through Ensemble Boosting

Luna Bloin-Wibe[1,*], Robin Noyelle[1,*], Vincent Humphrey[1,2], Urs Beyerle[1], Reto Knutti[1], and Erich Fischer[1]

[1]Institute for Atmospheric and Climate Science, ETH Zurich, Zurich, Switzerland
[2]Federal Office of Meteorology and Climatology MeteoSwiss
[*]These authors contributed equally to this work.

Correspondence: Luna Bloin-Wibe (luna.bloinwibe@env.ethz.ch)

Abstract.

With climate change, extremes such as heatwaves, heavy precipitation events, droughts and extreme fire weather have become more frequent in different regions of the world. It is therefore crucial to further their physical understanding, but due to their rarity in both observational and climate modeling samples, this remains challenging. For numerical simulations, one way to overcome this under-sampling problem is Ensemble Boosting, which uses perturbed initial conditions of extreme events in an existing reference climate model simulation to efficiently generate physically consistent trajectories of very rare extremes in climate models. However, it has not yet been possible to estimate the return periods of these simulations, since the conditional resampling alters the probabilistic link between the boosted simulations and the underlying original climate simulation they come from.

Here, we introduce a statistical framework to estimate return periods for these simulations by using probabilities conditional on the shared antecedent conditions between the reference and perturbed simulations. We validate this framework with a simple red-noise process and find the typical time scale at which one could expect to sample stronger extremes. This is then applied to simulations of the fully-coupled climate model CESM2: first for a pre-industrial control simulation, and then in present-day conditions, where, as an example, we estimate the return period of the record-shattering 2021 Pacific Northwest heatwave to be 2500 [2000-4000] years. Our evaluation of the method shows that return periods estimated from Ensemble Boosting are consistent with those of a 4000-year control simulation, while using approximately 6 times less computational resources. We thus outline the usage of Ensemble Boosting as an efficient tool for gaining statistical information on rare extremes. This could be valuable as a complement to existing storyline approaches, but also as an additional method of estimating return periods for real-life extreme events within a climate model context.

1 Introduction

Extreme weather events, or phenomena that occur at the tails of the climatological distribution, can have devastating impacts on ecosystems, human life, settlements, and infrastructure. In recent years, climate change has caused the frequency of extremes such as heatwaves, heavy precipitation events, drought and fire weather to increase (Ranasinghe et al., 2021; Seneviratne et al.,

2021). Additionally, the non-stationary climate that arises through climate change also means that record-shattering heatwaves are on the rise, which can pose challenges for adaptation, since communities are more vulnerable to extremes they have not yet witnessed (Fischer et al., 2021).

In particular, heatwaves have become not only more common, but also more intense and long-lived compared to the pre-industrial climate, a change which is projected to continue with climate change (Meehl and Tebaldi, 2004; Rahmstorf and Coumou, 2011; Thiery et al., 2021; Seneviratne et al., 2021). Several impactful summer heatwaves have been observed in the last decades: examples include the 2003 European heatwave, claiming approximately 70,000 lives (Robine et al., 2008), the month-long 2010 Russian heatwave (Otto et al., 2012) and the unprecedented Pacific North-West (PNW) heatwave of 2021, that broke observational records by more than 5 C (Bartusek et al., 2022; Malinina and Gillett, 2024; McKinnon and Simpson, 2022; Neal et al., 2022; Overland, 2021; Philip et al., 2022; Schumacher et al., 2022; White et al., 2023). Studying such heatwaves is societally relevant, due to their significant socio-economic and ecological impacts (Gourdji et al., 2013; Dunne et al., 2013) – in particular the high mortality tolls such heatwaves can incur (Robine et al., 2008; Vicedo-Cabrera et al., 2021).

Summer heatwaves in the mid-latitudes are associated with persistent anticyclonic flow anomalies that can give rise to prolonged anomalies of temperatures (Perkins, 2015; Horton et al., 2016; Barriopedro et al., 2023). These anomalies of temperature occur through three physical mechanisms: horizontal advection of warmer air from neighboring regions, adiabatic warming by subsidence within the anticyclone, and diabatic fluxes such as increased radiation due to the clear-sky conditions and sensible heat fluxes from the ground caused by reduced soil moisture (Pfahl and Wernli, 2012; Miralles et al., 2014). The relative importance of these mechanisms can vary greatly from one region to another (Röthlisberger and Papritz, 2023).

Despite the societal relevance of – and significant research on – heatwaves and their dynamics, their quantitative study remains challenging due to limited sample sizes. Extreme events are rare by definition, and it is therefore difficult to obtain accurate climatological results concerning their dynamics. Several methods have been developed to overcome this sampling challenge. For example, climate model Large Ensembles constitute a brute-force approach to the problem of sampling extreme events, and have been used to describe heatwave dynamics quantitatively (Suarez-Gutierrez et al., 2018; Schaller et al., 2018). However, the sampling density necessary to adequately represent events such as the PNW heatwave cannot be achieved even with these computationally expensive large ensembles (Fischer et al., 2023).

Another way in which extremes can be studied is through Extreme Value Theory (Coles et al., 2001). Extreme Value Theory is based on the existence of asymptotic results of the block-maxima and peak-over-thresholds distributions of any random variable. For example, generalized extreme value (GEV) distributions are often used to extrapolate probabilistic information on block maxima from the sample at hand (Philip et al., 2020; Cooley, 2013). However, such estimations can be problematic. Using a large ensemble of a climate model, Zeder et al. (2023), for example, showed that return period estimates of temperature extremes are systematically overestimated in short records. Additionally, while process-based covariates can provide some dynamical insight into heatwave drivers, this approach remains limited compared to the information available directly from fully-coupled climate models (Zeder and Fischer, 2023).

Therefore, there has been a recent push for methods that can generate climate simulations of extremes more efficiently than by producing ever larger climate model ensembles. One such tool is the application of rare event algorithms to climate models

(Wouters and Bouchet, 2016; Plotkin et al., 2019; Webber et al., 2019; Yiou and Jézéquel, 2020; Gessner et al., 2021; Finkel and O'Gorman, 2024). In particular, Ragone et al. (2018) used the Giardina-Kurchan-Tailleur-Lecomte (GKTL) (Giardinà et al., 2006; Giardina et al., 2011) algorithm to clone simulations that perform well with respect to a de ned score function to generate new, extreme, simulations. This process is repeated in intervals of a given resampling time; at each step, new simulations are given weights based on their parent score, which allows for a retracing of steps to calculate the probability of exceeding thresholds of time averaged quantities.

By design, the GKTL algorithm is well suited for extremes that persist in time, since time needs to elapse between the repeated cloning and evaluation steps. However, its use is limited when studying shorter extremes, like week-long heatwaves or daily precipitation extremes. In comparison, Ensemble Boosting, proposed by Gessner et al. (2021), is more suited for simulating short, very intense events. It consists in perturbing the initial conditions of extreme events selected from an already existing climate model simulation, thus ef ciently creating alternative, potentially substantially more extreme versions of a given extreme event in a parent simulation. However, in doing so, it breaks the probabilistic link between the resulting simulations and the climate model large ensemble they come from. Finding unconditional return periods directly through the boosted simulations is thus not straightforward. Therefore, Ensemble Boosting is typically used within a storyline framework (Fischer et al., 2023; Lüthi et al., 2024), an approach that seeks to complement probabilistic con dence statements with plausible, episodic information on representative case studies (Shepherd et al., 2018).

A key difference between the GKTL algorithm and Ensemble Boosting is that in the latter case, the perturbation is performed in anticipation of the extreme: this means that there is no guarantee that the boosted simulation will be at least as extreme as the parent event, thus breaking the necessary assumption for the weighted probability calculation. To overcome this, Finkel and O'Gorman (2024) showed that a chained conditional probability calculation based on the Subset Simulation framework (Au and Beck, 2001) could be used to estimate probabilities for ahead-of-time resampled simulations. However, this calculation is based on a resampling method that is methodologically distinct to Ensemble Boosting and can be dif cult to apply in typical simulation conditions of a climate model.

In this paper, we show how probabilities and return periods of extreme events found using the Ensemble Boosting resampling method can be estimated. In essence, we apply conditional probabilities — to calculate the unconditional probability of a boosted simulation, the unconditional probability of parent events selected for Ensemble Boosting is combined with the conditional probability of parents and boosted simulations given Ensemble Boosting. We additionally show, with a simple red-noise process, at which time scale Ensemble Boosting is expected to sample more extreme events.

The method is evaluated by generating extreme heatwaves for the region of the Paci c Northwest (45–52N, 115–123W) with the Community Earth System Model 2.1.2 (Danabasoglu et al., 2020), rst with pre-industrial (PI) control simulations as a proof of concept, and then with present anthropogenic forcing conditions. The 2021 PNW heatwave is used as a case study: by boosting analogue simulations under current climatic conditions, we attempt to nd the return period of this record-shattering heatwave within the model. We thus illustrate how we can both simulate the physical mechanisms and estimate the probability of very extreme, unseen events with a climate model.

The paper is organized as follows. Section 2 presents the theoretical framework of the Ensemble Boosting estimator for low probability events, showing that it is unbiased and can reduce the relative error compared to a naive estimator. In particular, we validate our approach with a simple 1D red-noise process. Section 3 presents the results for the pre-industrial control simulation with sensitivity tests, before applying the insights gained to the 2021 PNW heatwave in a present world context. In Section 4, the gain in computational resources, the choice of optimal parameter settings and the distribution of the very tail extremes are discussed. Finally, conclusions are presented in Section 5.

2 Methods

Symbol	Formal de nition	Description
PNW		Paci c North-West, region of interest in the study
TXx5d		Yearly summer max. of daily max. temperature anomalies with a 5-day running mean
Z500		Geopotential height at 500 hPa
N		Number of simulation years in the reference climate model simulation
N_{lead}		The number of lead times used to generate boosted simulations
N_{batch}		The number of boosted simulations per lead time
N_{parent}		The total number of simulations in the Parent Ensemble
N_b	$N_{lead} \quad N_{batch} \quad N_{parent}$	The total number of simulations in the Boosted Ensemble
T	$f T^n \; j \; n = 1 ; 2 ::: N \; g$	Set of TXx5d in the reference climate model simulation
T_b	$f T_b^m \; j \; m = 1 ; 2 ::: N_b g$	Set of TXx5d in the Boosted Ensemble
T_{ref}		User-chosen TXx5d threshold. If $T^n \quad T_{ref}$, it is selected to the Parent Ensemble
T_{ext}		TXx5d $\quad T_{ref}$ from the Boosted Ensemble
t		Lead time at which Ensemble Boosting is performed
AC_t^0	$f \; X_t \; j \; T_{t=0} \quad T_{ref} \; g$	Set of antecedent conditions at t of all members of the Parent Ensemble
AC_t	$f Y_t \; j \; 9 X_t \; 2 \; AC_t^0 ; k X_t \quad Y_t k \quad g$	Set of antecedent conditions that differ from AC_t^0 only by a perturbation at t
P		Theoretical probability
$\hat{p}$		Probability estimator
$E1$		The most extreme parent event in the 30-member Large Ensemble
$\tilde{E}1$		The most extreme parent event in the 100-member Large Ensemble

Table 1. De nitions and descriptions of important quantities used in this study.

2.1 Resampling low probability events

The return period of a heatwave reaching temperature T over a certain region is classically de ned as the inverse of the probability p_T of exceeding the temperature T. The latter can be estimated as the empirical frequency of occurrence:

$$\hat{p}_T = \frac{1}{N}\sum_{n=1}^{N} 1(T^n \quad T); \tag{1}$$

where $(T^n)_{1 \quad n \quad N}$ are N independent observations of the temperature T and 1 is the indicator function equals to 1 if $T^n \quad T$ and 0 otherwise. Since for each observation n, $E(1(T^n \quad T)) = p_T$,

$$E(\hat{p}_T) = p_T; \tag{2}$$

which makes this estimator unbiased. Furthermore, using the independence of each observation, its variance $V(\hat{p}_T)$ is $\frac{p_T(1 \quad p_T)}{N}$. Therefore, for $p_T \quad 1$, the relative error of this estimator $\hat{p}_T$, RE, is:

$$RE := \frac{\sqrt{V(\hat{p}_T)}}{E(\hat{p}_T)}, \quad \sqrt{\frac{1}{p_T N}}: \tag{3}$$

This shows that the relative error of the naive estimator increases as p_T decreases, i.e. when the temperature T becomes more extreme. In other words, the problem of obtaining precise climatological results for events reaching extreme T values of comes from under-sampling such events for a small number of observations N.

Ensemble Boosting addresses this problem by resampling the most extreme events of an already existing climate model simulation. In the following, we call these events the Parent Ensemble. The resampling is done by perturbing the antecedent conditions of the events inside the Parent Ensemble. Since the perturbations are of a relative amplitude of 10^{-18} and are only performed once, physically consistent simulations, hereby called the Boosted Ensemble, can be generated. All members of the Boosted Ensemble are samples from a distribution biased towards the upper tail of the distribution of T (see Appendix Figure A1) and, in particular, could sample events that are more intense than any present in the Parent Ensemble.

Figure 1 illustrates the Ensemble Boosting algorithm. A detailed description of the algorithm employed is presented below for temperature anomalies, but can be generalized to any quantity of interest that one wants to maximize or minimize:

1. A reference climate model simulation, spanning N years, is used to select the Parent Ensemble. Here, we consider the yearly summer maximum of daily maximum temperature anomalies with a running mean of 5 days (hereby denoted by TXx5d), forming a set of temperatures $T := \{T^n \mid n = 1, 2, \ldots, N\}$, with n indicating the different simulation years. All years whose T^n exceeds a user-chosen temperature threshold T_{thr} are selected to form the Parent Ensemble. The total size of this ensemble is denoted by $N_{parent} \quad N$. An illustration of the reference climate model simulation can be seen in Figure 1a, and an example of an event in the Parent Ensemble can be seen in Figure 1b.

2. Since the objective of Ensemble Boosting is to generate a heatwave that is similar to, but substantially more intense than the parent event, each parent event in the Parent Ensemble is perturbed ahead of the parent event's peak. The number of days between the perturbation and the parent peak is called lead time and denoted by τ. Because of the chaotic nature

Figure 1. Illustration of the Ensemble Boosting algorithm for heatwaves1) The reference climate model simulation spanning the period 1801–1850. Each dot represents the yearly summer maximum of daily maximum temperature anomalies with a running mean of 5 days (TXx5d). 2) One parent event of the Parent Ensemble: the heatwave itself is represented as a function of lead time, de ned here as the number of days before the parent event's heatwave peak (solid black vertical line). The y-axis is the 5-day running mean daily maximum temperature anomaly (Tx5d anomaly). 3) Boosted simulations (orange) of the parent event shown in 2), perturbed at a lead time of (a) 16 days, (b) 12 days and (c) 7 days. The perturbation lead time is highlighted by a black dashed vertical line. At each lead time, a "batch" of 100 boosted simulations are generated.

of the model, the deviations between the boosted simulations and the parent event will initially grow exponentially with time (Lorenz, 2006). Figure 1d, e show boosted simulations that have been perturbed at different lead times. Here, we see how the lead time in uences the temperatures in the Boosted Ensemble: simulations perturbed 16 days in Figure 1c have diverged for too long, thus reverting back to the underlying climatology. Therefore, only a few boosted simulations reach the parent TXx5d. Simulations perturbed 7 days, on the other hand, as seen in Figure 1e, have not diverged enough for any boosted simulation to deviate substantially from their parent. Figure 1d, 12 days, illustrates an ideal lead time, i.e. where the spread is neither too small nor too large, and a large portion of boosted simulations exceed the parent TXx5d. We propose an estimation of this "optimal lead time" in the case of a red-noise process below. The difference between the climatological distribution and the distribution of the boosted simulations at this lead time is shown in Appendix Figure A1.

3. The perturbations are performed on the 3D speci c humidity Q at each grid cell, to obtain the perturbed speci c humidity eld Q^m for each boosted member m:

$$Q^m_{i;j;k} = Q^m_{i;j;k} \quad 1 + 10 \quad {}^{13} R^m_{i;j;k} \quad ; \tag{4}$$

for a given boosted simulation m and a given grid cell $i; j; k$. $R^m_{i;j;k}$ is a random term drawn from a uniform distribution between $\frac{1}{2}$ and $\frac{1}{2}$. Each newly generated offspring simulation is then run for 21 days, and the maximum TXx^m_b, TXx5d, is considered. While a systematic test has not been implemented, we do not expect the choice of variable to in uence results, since the perturbation stays within numerical noise limits. Both speci c humidity and temperature have been used in previous studies, with comparable results (Gessner et al. (2021), Gessner et al. (2022), Gessner et al. (2023) Fischer et al. (2023)).

4. For each lead time and each parent, new simulations are generated in batches of size N_{batch}, corresponding to different realizations of the perturbed eld $(Q^m)_{1 \quad m \quad N_{batch}}$. The total number of simulations in the Boosted Ensemble N_b is therefore:

$$N_b = N_{lead} \quad N_{batch} \quad N_{parent} \; ; \tag{5}$$

where N_{lead} is the number of lead times where perturbations are performed.

Computing the probability of exceeding an extreme temperature T naively from simulations of the Boosted Ensemble using Equation 1 leads to a probability estimation conditional on the initial conditions at lead time l of the parent events. To recover the unconditional, i.e. climatological, probability we use the framework of Subset Simulation (Au and Beck (2001)), where an iterative chain of conditional probabilities is used to estimate the probability of simulated low-probability events. Here, we illustrate how this framework can be used with a deterministic climate model, with perturbations ahead of time as in Finkel and O'Gorman (2024), and for the Ensemble Boosting setup, which generates batches of boosted simulations, but here stops after one perturbation iteration.

The probability of exceeding the threshold temperature T_{ref} — used to select events for the Parent Ensemble — can be estimated using the naive estimator of the reference climate model simulation (Equation 1), since there are, by de nition, events in the reference climate model simulation that exceed T_{ref}. This also means that all parent events have in common that the dynamical conditions of the climate model system in the days leading up to the peak were such that the temperature T_{ref} was reached.

One can thus de ne a set that includes the antecedent conditions at lead time t for all the parent events:

$$AC_t^0 := \{ X_t \mid T_{t=0} \geq T_{ref} \}; \tag{6}$$

where X_t is the state vector of the climate model at lead time t, which encompasses all the degrees of freedom of the system. Here, X_t is associated with a temperature $T_{t=0} := T(X_{t=0})$ in the location of interest. The set AC_t^0 is imperfectly sampled in a simulation with a nite length, but we assume that the antecedent conditions of the selected parent events still represent a good enough sample. This is a key assumption and it will be further discussed in the Results and Discussion sections. The exact conditions constituting a representative sample are unknown, since we do not know the function giving the probability to reach an extreme given the current state — the so-called committor function (see Miloshevich et al., 2023). Additionally, if we did, it may not be transferrable across different regions. Instead, temperature at $t_0 = 0$ is used as a rst order approximation to determine the right antecedent conditions (at lead time $t = t_0$) for a heatwave — if the antecedent conditions produced a heatwave, it could potentially produce a more extreme one.

Since the antecedent conditions of the boosted simulations only differ from their parent by the value of the perturbation at lead time t, they are samples of the set $AC_t := \{ Y_t \mid \exists X_t \in AC_t^0; k X_t - Y_t k \leq \epsilon \}$, for a distance $k \cdot k$ and ϵ the absolute value of the perturbation. By de nition, $AC_t^0 \subseteq AC_t$. An illustration of these two sets, and how they relate to each other, can be seen in Figure 2. Note that for the boosted simulations, T_0^m can either be in AC_t^0 (and by extension, in AC_t) or only in AC_t, since there is no guarantee that they will exceed T_{ref} at $t = 0$.

The probability of any temperature $T_{ext} \geq T_{ref}$ from the Boosted Ensemble can thus be estimated by combining the conditional probability of T_{ref} and T_{ext}. These probabilities can be stated using the conditional probability de nition as:

$$P(T_{t=0} \geq T_{ref} \cap AC_t) = P(T_{t=0} \geq T_{ref} \mid AC_t) P(AC_t) \tag{7}$$

$$P(T_{t=0} \geq T_{ext} \cap AC_t) = P(T_{t=0} \geq T_{ext} \mid AC_t) P(AC_t): \tag{8}$$

Since $T_{ext} \geq T_{ref}$ and $AC_t^0 \subseteq AC_t$, the presence of antecedent conditions AC_t at time $t = 0$ and getting temperatures above either T_{ref} or T_{ext} at time $t = 0$ overlap perfectly. Therefore $P(T_{t=0} \geq T_{ref} \cap AC_t)$ simpli es to $P(T_{t=0} \geq T_{ref})$. An equivalent simpli cation can be performed for $P(T_{t=0} \geq T_{ext})$. The subscript $t = 0$ indicates that this is the probability at the day of the parent event's peak. Since all values considered in this study are summer temperature anomalies, thus not presenting any seasonality changes with respect to each other, we will assume that this probability is the same as the climatological probability $P(T \geq T_{ext})$, i.e. for any time within the summers of the reference climate model simulation.

From Equations 7 and 8, we can derive the probability of T_{ext}:

Figure 2. Schematic illustration of the antecedent condition set AC_t and the stricter AC_t^0 AC_t. The dark blue line shows the delineation between AC_t and AC_t^0, i.e. the condition that at $t = 0$, $T = T_{ref}$. Only simulations that exceed this requirement (the Parent Ensemble) are selected from the reference climate model simulation, and are shown in blue diamonds. Other events in the reference climate model simulation are not depicted. The boosted simulations are shown as orange dots. Since the boosting perturbation is performed ahead of the parent peak, there is no a priori guarantee that at $t = 0$, T_b^m T_{ref}, and the orange dots can therefore be on either side of the delineation.

$$P(T \quad T_{ext}) = P(T \quad T_{ref}) \frac{P(T \quad T_{ext} \mid AC_t)}{P(T \quad T_{ref} \mid AC_t)}. \tag{9}$$

Here, $P(T \quad T_{ref})$ is the probability of reaching T_{ref} in the reference climate model simulation, $P(T \quad T_{ref} \mid AC_t)$ is the probability of reaching T_{ref} in the Boosted Ensemble, and $P(T \quad T_{ext} \mid AC_t)$ is the probability of reaching T_{ext} in the Boosted Ensemble. In other words, the unconditional probability of exceeding the threshold is equal to the product of the probability of exceeding the threshold given one has entered the set, times the probability of entering the set. Note that since the condition $T_{ext} \quad T_{ref}$ is necessary to derive Equation 9, we can only use it to nd probabilities for boosted simulations where T_{ref} is exceeded.

All terms in this equation can be estimated with the climate model simulations at hand: the probability of reaching T_{ref} can be approximated to its frequency of occurrence in the reference climate model simulation, while the probability of reaching T_{ref} and T_{ext} given AC_t can be approximated to their respective frequencies in the Boosted Ensemble. Formally this can be written as:

$$\hat{p}_{T \geq T_{ext}} := \hat{p}_{T \geq T_{ref}} \frac{\hat{p}_{T \geq T_{ext} \mid AC_t}}{\hat{p}_{T \geq T_{ref} \mid AC_t}} = \frac{1}{N} \sum_{n=1}^{N} 1(T^n \geq T_{ref}) \left(\frac{\frac{1}{N_b} \sum_{m=1}^{N_b} 1(T_b^m \geq T_{ext})}{\frac{1}{N_b} \sum_{m=1}^{N_b} 1(T_b^m \geq T_{ref})} \right). \tag{10}$$

The estimator $\hat{p}_{T \geq T_{ext}}$ will hereby be referred to as the *boosting estimator*. As shown in Noyelle (2024) and derived in Appendix Section A, this estimator is unbiased, if the following assumptions are made:

1. the term $\hat{p}_{T \geq T_{ref}}$ is independent from the ratio $\frac{\hat{p}_{T \geq T_{ext} \mid AC_t}}{\hat{p}_{T \geq T_{ref} \mid AC_t}}$. This means that the probability of T_{ref} is independent from how much more likely it is to reach T_{ref} than T_{ext} within the Boosted Ensemble.

2. the indicator variables $(1(T_b^m \geq T_{ext}))_{1 \leq m \leq N_b}$ and $(1(T_b^m \geq T_{ref}))_{1 \leq m \leq N_b}$ need to be independent.

These assumptions will be discussed in the discussion section 4.3. Note that these assumptions do not require that T_{ref} and T_{ext} occur independently of one another, but rather explore the relationship between T_{ref} and T_{ext} within the boosted ensemble, and their link to T_{ref}.

The choice of an intermediate temperature threshold to create a link between the reference climate model simulation and the Boosted Ensemble has been studied, albeit under different conditions, in Finkel and O'Gorman (2024). However, beyond the higher complexity of the climate model used here for Ensemble Boosting compared to that of the Lorenz 96 simulator used in Finkel and O'Gorman (2024), the most important methodological difference lies in the process of generating perturbed offspring. While Finkel and O'Gorman (2024) boost one parent with a batch of 1 repeatedly until reaching the desired extremes, calculating conditional probabilities at each step, we perturb larger batches for ranges of lead times (as detailed above) only once. The former approach would lead to a more targeted result, with fewer unused simulations. However, it would also take substantially more time to perform all computations, since there would be no possibility of parallelization: given our current computational capacity (simulating approximately 1.1 climate model years in 24 node hours, with a capacity of up to 4 nodes), running the same 12 000 simulations generated for this study with this approach would take approximately 1.5 node years. Running large batches allows us to complete such setups within a few weeks. There is thus a trade-off between minimizing compute time using very effective sampling vs. minimizing wall-clock time by running more parallel simulations.

2.2 Theoretical comparison to the naive estimator

In order for the boosting estimator to be useful beyond $\hat{p}_{T \geq T_{ref}}$, it needs to have lower errors than the naive estimator. In Appendix Section A, the variance and relative error of the boosting estimator are calculated, and can be shown to depend on the number of simulation years in the reference climate model simulation N, the total number of simulations in the Boosted Ensemble N_b, and the three probability terms $P(T \geq T_{ref})$, $P(T \geq T_{ext} \mid AC_t)$ and $P(T \geq T_{ref} \mid AC_t)$.

Therefore, in order to calculate the theoretical relative error of estimating $P(T \geq T_{ext})$, we need to estimate the above unknowns. While the first three terms N, N_b and $P(T \geq T_{ref})$, are parameters that can be set by the experimenter, $P(T \geq T_{ext} \mid AC_t)$ and $P(T \geq T_{ref} \mid AC_t)$ can only be calculated once the boosting experiment is performed, and will have to

Figure 3. Theoretical relative error of the boosting estimator with $N = 50 \times 100$ days and (solid orange line) $N_b = 100 \times 21$ days, (dashed orange line) $N_b = 500 \times 21$ days and (dotted orange line) $N_b = 3000 \times 21$ days. Theoretical relative errors computed for (a) $\hat{p}_{T > T_{ref} | AC_t} = 0.75$, and (b) $\hat{p}_{T > T_{ref} | AC_t} = 0.3$. Since a boosted simulation is less computationally expensive to run (see text), the naive estimator with a equivalent computational resource for each configuration of the boosting estimator is represented by a blue solid, dashed and dotted line, respectively. The relative error of the naive estimator with (solid line) $N = 1000 \times 100$ days and (dashed line) $N = 50 \times 100$ days are shown in black.

be approximated based on empirical evidence. Furthermore, T_{ref}, T_{ref} and T_{ext} could exceed any T_b^m from the Boosted Ensemble. Therefore, $P(T > T_{ext} | AC_t) \in [0; P(T > T_{ref} | AC_t)]$. For illustration purposes, we choose values of 0.75 and 0.3 for the estimation of $P(T > T_{ref} | AC_t)$, which amounts to cases where either 75% or 30% of the boosted simulations exceed T_{ref}. These values were chosen because they correspond to the typical value we find in practice (see Results section 3.2).

The evolution of the theoretical relative error of the boosting estimator $P(T > T_{ext})$, for different values of the above mentioned parameters, can be seen in Figure 3. Here, three different configurations of the boosting estimator are compared to the errors of the naive estimator. Since we need to run a sample of non-boosted parents before generating boosted samples, the total cost of the boosted simulations takes this into account ($N + N_b = 50 \times 100 + 500 \times 21$ days, for a parent ensemble of size 50, where each simulation is run for 100 days, and a boosted ensemble of size 500, where each simulation is run for 21 days). To directly compare these results with the naive estimator, we generate a non-boosted sample N is equivalent, in terms of computational resources, to those of each boosting configuration. For example, $N = 105 \times 100$ days is equivalent to $N + N_b = 50 \times 100 + 500 \times 21$ days, since a boosted simulation only needs to be run for 21 days, while the reference climate model simulation needs approximately 100 days to generate a full summer.

An important result of this computation is that the relative errors of the boosting estimators are constrained by that of $P(T \quad T_{ref})$: any estimate of $P(T \quad T_{ext})$ will have errors equal to or higher than $P(T \quad T_{ref})$. An accurate estimate of $P(T \quad T_{ref})$ is therefore necessary for robust results. This advocates for using either a longer test slice as the reference climate model simulation or a less extreme (and thus better sampled) T_{ref}.

Nonetheless, the boosting estimator can reduce the relative error compared to that of the naive estimator. Firstly, in all boosting con gurations, relative errors are smaller than for the naive estimator with $N = 50 \quad 100$ days. Secondly, each boosting estimator reduces errors compared to its equivalently expensive naive estimator for probabilities under a certain threshold value of $p_{T \quad T_{ext}}$, which depends on $N + N_b$. Finally, the boosting estimator with $N = 50 \quad 100$, $N_b = 3000 \quad 21$ days reduces errors compared to the much more expensive naive estimator with $N = 4000 \quad 100$ days in both Figure 3 a and b, for a low enough probability $p_{T \quad T_{ext}}$.

Finally, Figure 3 also shows that for both N_b and $P(T \quad T_{ref} \mid AC_t)$, the relative error decreases as these parameters increase. There is therefore no optimal number of boosted simulations: N_b is the larger the better. Finding the optimal $P(T \quad T_{ref} \mid AC_t)$ is however less trivial because it creates trade-offs. These will be further discussed in Results section 3.2.

2.3 Validation with an Ornstein-Uhlenbeck process

Before estimating return periods in a fully-coupled climate model, a validation and exploration of parameter settings is performed with a simple Ornstein-Uhlenbeck process, also called red-noise process. The objective of this validation, beyond checking that the estimator yields correct results, is to better quantify the uncertainties surrounding the boosting estimator, by separating the effect of parameter settings and random variability. Due to the lack of computational constraints in running the Ornstein-Uhlenbeck process, larger samples of simulations can be generated, which strengthens con dence in the conclusions drawn.

The evolution of this process obeys the following stochastic equation:

$$dX(t) = \quad X dt \; + \; dW(t); \tag{11}$$

where the statistical parameters, $\quad$ are equal to 1 here, and $dW(t) \quad N \quad (0; dt)$ is a Wiener process. It is composed of two terms: the rst one, $X dt$, models a deterministic drift towards 0, while the second one, $dW(t)$, generates randomness that simulates natural variability. This creates the characteristic mean-reverting evolution of the Ornstein-Uhlenbeck process.

We simulate 10^6 parallel simulations of this process that span $100 \quad$, with $= \frac{1}{\quad}$ the de-correlation time of the process. This assures independence from each simulation's initial value, which is sampled from the stationary distribution of the Orstein-Uhlenbeck process. The set of maximum values, or events, found for each of these parallel simulations are calculated to act as a ground truth. From this ground truth, we sample $N = 1000$ events from which we select the parent ensemble. The boosted ensembles with varying N_{parent} and N_{batch} are subsequently generated. Since the dynamics of spread in a red-noise process is not directly comparable to that of a fully-coupled climate model, we keep $N_{lead} = 1$, and perturb at a lead time $t = 0:2 \quad$. Other lead times were tried, with no substantial change in results (not shown). Each simulation is then run until it is no longer

Figure 4. Estimating return periods from boosted simulations in an Ornstein-Uhlenbeck process. (a) Evolution in time of one parent (blue) and its boosted simulations (orange). The mean ± 3 standard deviations of the boosted simulations is shaded in orange. Additionally, the theoretical mean (dashed line) and mean ± 3 standard deviations (solid line), as derived in Appendix Section B, are shown in black. The time at which the maximum theoretical boosted intensity is reached is highlighted by black dotted lines.
Estimated return periods from boosted simulations for (b) $N_{parent} = 10$, and $N_{batch} = 10$, (c) $N_{parent} = 100$, and $N_{batch} = 10$, (d) $N_{parent} = 10$, and $N_{batch} = 100$, and (e) $N_{parent} = 100$, and $N_{batch} = 100$ are shown in orange: the median value, sampled from 1000 boosting experiments, is shown with a solid line, while the 95% con dence interval is delineated by dotted lines. In each panel, the ground truth is shown in black, while the 1000 repeated samples of the simulations used to select the parent ensemble are shown in blue (the median value as a solid line, and the 95% con dence interval is shaded).

correlated to its parent (time τ), and its maximum value is calculated. This process is repeated 1000 times to estimate uncertainties.

The evolution in time of one parent simulation, along with the mean ± 3 standard deviations of 1000 boosted simulations is shown in Figure 4. Since the Ornstein-Uhlenbeck process employed here is fully stochastic, the typical spread of boosted simulations from their parent does indeed differ from the evolution of boosted simulations in a fully-coupled climate model as seen in Figure 1. However, the spread between members, illustrated by the range of the ± 3 standard deviations con dence interval, grows and saturates faster than the memory of the parent intensity, shown by the median of the boosted simulations, fades and reverts back to zero. This opens a window of time when one can sample events more intense as the parent x_0. As derived theoretically in Appendix Section B, for the Orstein-Uhlenbeck process, if one wants to sample anomalies of size k, with $k > 0$ being a function of the number of members in the boosted ensemble, then the time where the maxima are expected to happen (i.e. an approximation of the "optimal lead time") is:

$$t = \frac{1}{2}\ln\left(1 + k^2 \frac{{}^2/2}{x_0^2}\right): \tag{12}$$

As a sanity check, this equation can be applied to our climate model setup (see Subsection 2.4). When considering the evolution of temperature in summer, one would typically have $\tau \sim 10$ d (the de-correlation time scale in the atmosphere), $= \frac{p}{2} \sim 3.5$ K (the climatological standard deviation of summer 5-day rolling average of daily maximum temperature anomalies over the region of interest), $x_0 \sim 10$ K (the typical value of T_{ref}), and $k \sim 3$ (the maximum number of standard deviations one can expect to sample using $N_B \sim 100$ to 1000 boosted members per parent). This leads t to 15 d, with expected maximum event magnitudes around 4 climatological standard deviations, i.e. temperature anomalies of 14 K.

Figure 4b-e shows return period estimates in the boosted ensemble, calculated with the boosting estimator for combinations of $N_{parent} = 10; 100$ and $N_{batch} = 10; 100$. In all panels, the return period of the ground truth set, estimated with Equation 1, (in black) and sample sets from which we select each parent ensemble (in blue) are shown for comparison. Firstly, we see that the boosting estimator is indeed unbiased – all median return period estimates using the boosting estimator follow the ground truth, regardless of the con gurations of N_{parent} and N_{batch}. Secondly, in all con gurations of N_{parent} and N_{batch}, higher return periods, up to more than an order of magnitude, are estimated in the boosted ensemble than in the parent ensemble.

Additionally, while median return period estimates of the sample sets from which we select the parent ensemble also follow the ground truth (as expected, since both are estimated using Equation 1), the uncertainty of the return periods estimated from the boosted simulations seem to increase more slowly than for the non-boosted simulations in the sample set. Indeed, the 95% con dence interval around the return periods of the boosted simulations are narrower than those of the sample set for return periods above 10^2. As predicted from the theoretical relative error calculation (see Section 2.2), this reduction of uncertainty is more pronounced for a larger N_{batch} (Panels d,e) or N_{parent} (Panels c,e), with the smallest uncertainty range found in Panel e, where the total computational cost is largest (10 000 simulations). This highlights that while sparse sampling of either the parent or the boosted ensemble can increase uncertainty and relative error with respect to the ground truth, it does not bias the return period estimation.

Finally, comparing Panels c and d allows for evaluating how N_{parent} and N_{batch} affect the boosted simulations and their subsequent return period estimation. At the same computational cost as Panel c, Panel d presents boosted simulations with higher return levels and similar, or at times smaller, levels of uncertainty. This is because $P(T \geq T_{ref})$ here is much higher (0.01 compared to 0.1 for Panel c), thus increasing the chances of events in the boosted ensemble with larger return periods.

By exploring different configurations of the relevant boosting parameters N_{parent} and N_{batch}, this evaluation thus strengthens confidence in estimating return periods using the boosting estimator. This confidence is underscored by a more extensive uncertainty sampling of both the parent and boosted ensemble than what would be possible in a fully-coupled climate model. However, it is important to note that the Ornstein-Uhlenbeck process is one-dimensional, while a state-of-the art climate model has millions of degrees of freedom. Therefore, the transferability of results found here are limited, in particular for the size of the parent ensemble necessary to sample AC_t well.

2.4 Experimental setup of Ensemble Boosting in a fully-coupled climate model

To show how the theoretical properties of the boosting estimator apply in a climate model context, beyond a simple Ornstein-Uhlenbeck process, we generate simulations with the state-of-the-art fully coupled Community Earth System Model 2.1.2 (Danabasoglu et al., 2020). We seek to estimate the return levels of the yearly maximum of 5-day rolling average of daily maximum temperature anomalies (TXx5d) spatially averaged over the region of the PNW heatwave (45–52°N, 115–123°W), corresponding to the region used by Fischer et al. (2023).

First, we test our boosting estimator in a pre-industrial setting. A 4000-year long pre-industrial control run is used to act as a control period, while two 50-year time slices of this simulation, test slice 1 (1801–1850) and test slice 2 (1851–1900), act as our reference climate model simulations. While these ranges are adjacent to one another, the starting point of the total time range is selected randomly within the control run. Additionally, since the time is only referenced as time since the start of the simulation, the years do not bear any meaning or relation to real-world weather at that time. Two separate test slices are used to increase the robustness of the results. Since they only span 50 years, extremes (i.e. events with a return period longer than 50 years) will be scarce or absent and their return period estimates highly uncertain. The length of the time slice is selected to reflect typical timescales of available historical records. While this can serve as a comparison to historical extreme event attribution studies, this time scale might present highly uncertain estimations of T_{ref}, in particular due to the limited sample and the long term temperature variability effects. All events in the Parent Ensemble, selected according to the boosting algorithm above, are boosted according to the algorithm detailed above, and their return periods are calculated using Equation 10. This is done independently for both test slices. The results can then be directly compared against the more robust return periods found in the control period. Additionally, we fit stationary GEV distributions using both the maxima of the control period and of the test slices. This is done to give an estimate of return periods beyond these observational records with a standard method from Extreme Value Theory. A bootstrapping approach is used to calculate 95% confidence intervals, fitting GEV distributions to 1000 resamples drawn randomly from the test slices and the control period.

In each test slice, the five most extreme years are selected to form the Parent Ensemble, which corresponds to T_{ref} taking such that $P(T \geq T_{ref}) = 0.1$. This value is chosen to balance rareness in the Parent Ensemble and the ability to robustly estimate

$P(T \geq T_{ref})$ in the test slice. Changing the threshold of selection is equivalent to either changing T_{ref} and keeping the sample size equal (which means changing the number of parents) or keeping T_{ref} but changing the sample size. The former will be evaluated in the discussion section, and the latter in the following subsection. In order to sample the uncertainty of $P(T \geq T_{ref})$, this value is estimated by bootstrapping the relevant test slice sample. These are then subsequently boosted for all lead times in the range of 7 to 18 days, with batches of 100 simulations per lead time and per parent event. This leads to a total of 6000 boosted simulations for each test slice. The maximum 5-day daily maximum temperature (Tx5d) anomaly over the 21 days following perturbation is assessed, so that the block maximum approach stays the same as for the Parent Ensemble. The length of 21 days is taken to correspond roughly to the saturation time of the boosting algorithm, i.e. when the growing divergence between boosted simulations have saturated to that of the reference climate model simulation. This is in line with the time scale of classical weather predictability of around 10 days (Krishnamurthy, 2019). The exact length is determined through empirical trial and error, and will be discussed in Results section 3.2.

In order to directly compare return periods calculated using the boosting estimator, the TXx5d need to reasonably fulfill the domain convergence conditions of GEV theory. While 21 days is not as long as the approximately 100 days of a full summer, we postulate that the boosted simulations comply reasonably well with EVT requirements for the following reasons: Firstly, since T_{ref} is already a rare summer maximum, the boosted simulations that exceed it are likely to also be the summer maximum. Indeed, for the simulations that do exceed T_{ref}, only 4-5% have a higher temperature in the following 60 days. Secondly, the parent heatwaves do not all occur in the beginning of the summer. Since the boosted simulation shares the trajectory of its parent until it is perturbed, the boosted simulations are generally longer than just 21 days.

A bootstrapping procedure is also performed to calculate a confidence interval around the estimated return period from the boosting estimator. However, since the number of boosted simulations where $T_{ext} \geq T_{ref}$ is not always the same in random samples of the Boosted Ensemble (since some boosted simulations do not exceed T_{ref}), the bootstrapping proceeds as follows: for each $T_{ext} \geq T_{ref}$ in the Boosted Ensemble, 1000 random samples of the Boosted Ensemble are generated to calculate $P(T \geq T_{ref} \mid AC_t)$ and $P(T \geq T_{ext} \mid AC_t)$, and thus estimate $P(T \geq T_{ext})$. Note that since we do not change the experimental sample when calculating the confidence interval (e.g. selecting and boosting a different parent ensemble), this confidence interval may underestimate certain sources of uncertainty such as slow modes of variability of the climate system.

Second, we boost simulations from the CESM2.1.2 climate model in present-like conditions (2005-2035, SSP3-7.0 after 2015) to estimate the return period of the record-shattering 2021 Pacific North-West heatwave. A 30-member Large Ensemble is used as the reference climate model simulation, leading to $N_{ref} = 30 \cdot 31 = 930$ years. From this initial simulation, 7 among the 13 most extreme events are selected to form the Parent Ensemble, which is equivalent to a $P(T \geq T_{ref}) = \frac{13}{930} = 0.014$.

Due to the presence of anthropogenic forcings in the reference climate model simulation from 2005 to 2035, the non-stationarity in the underlying statistical distribution needs to be accounted for. This is done by linearly de-trending the TXx5d time series: for each year and each member of the Large Ensemble, a 100-member Large Ensemble of the same model (Rodgers et al., 2021) is used to produce a day-of-year mean across members and a three-year window (the year in question, and one year before and after) around each year. The three-year window was chosen to create a larger sample of similar years, without adding the climate change signal present over longer time scales. While simple de-trending could produce artifacts in the

distributional tail, the forcings in this historical and near-future sample is small enough for the correction to be considered a reasonable approximation.

The 100-member Large Ensemble spans the same time range (31 years) and is also corrected for non-stationarity in the same way. However, this publicly available data set is not locally bit-by-bit reproducible, which is necessary to generate boosted simulations. It will therefore act as a separate, larger data set that can provide return periods that are more precise than those of the 30-member Large Ensemble, since here, $N = 100 \cdot 31 = 3100$ years.

Note that the 7 selected events for present-like conditions are not, contrary to the pre-industrial control case, those with the absolute highest TXx5d, but rather 7 among the top 13 events. This is because the correction for non-stationarity was performed after the selection process, which was determined in a previous study (Fischer et al., 2023). In the context of the boosting estimator, this simply corresponds to a more sparse sampling of $\Delta C_{g,t}$ than what is maximally possible with the climate model simulation at hand: one could imagine a reference climate model simulation with only these 7 events present.

To calculate the TXx5d of the 2021 PNW heatwave, a detrended time series of the ERA5 reanalysis data set of the ECMWF (Hersbach et al. (2020)), regridded to t the CESM2.1.2 model grid, is used. The return period of this return level is then calculated using the boosting estimator, and, for comparison, the GEV ts of the 30- and 100-member Large Ensembles. A median (percentile) return period estimate is written as ∞ when the median (percentile) probability is 0.

3 Results

3.1 PI-control runs

Figure 5 shows return period estimates for TXx5d in the Boosted Ensemble, calculated with the boosting estimator, for test slice 1 (Panels a,b) and test slice 2 (Panels c,d). For comparison, return period estimates calculated with the naive estimator are shown for the control period, and a GEV t of the test slice itself provides an extrapolation outside the $t$ting period. In order to increase diversity and robustness among the boosted simulations, results are calculated by pooling together perturbation lead times from 18 to 13 days in Panels a,c, and 12 to 7 days in Panels b,d. This categorization remains somewhat arbitrary, however; results calculated lead time by lead time for test slices 1 and 2 can therefore be seen in Appendix Figures A2 and A3, respectively.

In Figure 5, the results found using the boosting estimator stand in stark contrast to those of the non-boosted test slices. First, the maximum TXx5d sampled through boosting substantially exceed those of the test slices, by up to 2.9 °C, and even reaches the maximum TXx5d of the control period (Figure 5b). Second, return period estimates of the simulations in the two test slices deviate more from the control period than those of the Boosted Ensemble, in particular for Test slice 1. This follows the theoretical results presented in Figure 3: for $N_g = 3000$ years, any con $g$uration of the boosting estimator should lead to return period estimates with less relative error than for the naive estimator with $N = 50$ years. The GEV distribution $t$ted to the test slices is also error-prone; for return periods up to $\frac{1}{P(T = T_{ref})}$, estimates calculated with the boosting estimator in all con $g$urations (Panels a, b, c, d) follow the control period better than the median GEV ts of the test slices. It is also worth noting that the con $d$ence interval of the GEV t of both test slices is much wider than that of the boosting estimator; in

Figure 5. Return periods of TXx5d in the PNW region estimated with the boosting estimator under stationary climate conditions. Estimated return periods with the boosting algorithm for parent events selected from test slice 1 (1801–1850) and (c,d) from test slice 2 (1851–1900) of the pre-industrial control simulation are shown in orange. Perturbation lead times are pooled together (a,c) from 18 to 13 days and (b,d) 12 to 7 days. In each panel, the TXx5d of the 4000 years of the pre-industrial control simulation are shown in black. The TXx5d of the two 50-year test slices are shown in blue. The five selected parent events in each test slice are highlighted in diamonds. The horizontal dashed orange line represents the reference $TXx5d_{ref}$ in each test slice. For the return periods of the control and test slice simulations, a GEV law is fitted and the estimated return period is shown (solid line) with a bootstrap 95% confidence interval (shaded). For the boosted simulations, the shaded area shows the bootstrap 95% confidence interval (see Methods).

particular, the upper bound of the confidence interval of test slice 2 is unbounded. These deviations from the control period are to be expected given the short climate model record (N_{rec} = 50 years) – both Zeder et al. (2023) and Noyelle et al. (2024a) have shown that GEV fits from short records are prone to systematic biases.

Furthermore, Figure 5 shows that return period estimates, calculated with the boosting estimator for lead times from 18 to 13 days in test slice 1, follow the estimate using the control period for all T_{ref} present in this sample: both median boosting estimates and the con dence interval overlap with the control period remarkably well. Results from test slice 2, for lead times from 18 to 13 days (Figure 5), indicate that estimates using the boosting estimator also can deviate from the control period, although they remain within an order of magnitude compared to the control period estimate. A reason for this deviation might be that the test slices span only 50 years, and only 5 parent events were selected for boosting. Therefore, it is possible that the Parent Ensemble does not sample AC_t suf ciently well, and that the return period estimates calculated are actually conditional on a particular feature of long term natural variability present in test slice 2 only. In Appendix Figure A4, the return periods calculated by combining the two test slices into one 100-year long time series are shown. Here, the return periods again follow the control period very well.

For shorter lead times, between 12 and 7 days, return period estimates deviate from the control period con dence interval already at around 30–40 years for simulations from both test slices (Figure 5b, d). This is likely because the short lead time leads to simulations that are too constrained by the maximum TXx5d of the parent events, since they do not have enough time to deviate enough from their parents. In Appendix Figures A2 and A3, it can be seen that this deviation is gradual, that appears around 13 days for Test slice 1, although for Test slice 2, the picture is less clear — some deviation is present already at 18 days. In particular, we observe "step-like" behavior of the boosted return periods around the values of the parent events for shorter lead times (Appendix Figure A2, j, l), meaning that the TXx5d of the boosted simulations are very similar to that of their parent, forming return levels that resemble a step function. If the boosted simulations are too close to their parents, this additionally means that there is less independence between boosted simulations, thus breaking one of the assumptions made the methods to show the unbiasedness of the boosting estimator. This could also contribute to why we have such a discrepancy for short lead times.

It is worth noting that the increase in estimation uncertainty, illustrated by a widening of the con dence interval as return periods increase, shows that the most extreme events are estimated less precisely. Additionally, the con dence intervals should be interpreted somewhat cautiously, since they are computed through bootstrapping. This means they are fundamentally limited by the sample at hand, and not representative of all uncertainty factors such as potential multi-decadal variability beyond the time scales sampled here.

3.2 Sensitivity tests: theoretical assumptions and parameter choices

The difference in results between longer and shorter lead times, seen in Figure 5, as well as the differences in relative error depending on N_b and $P(T \quad T_{ref} \mid AC_t)$ seen in Figure 3, show that the parameter choices and assumptions made can in uence the quality of the estimation obtained. This warrants a deeper analysis of these choices and assumptions.

First, the effects of the lead time chosen when boosting are assessed beyond the separation of short and long lead times in Figure 5. Figure 6 shows the TXx5d of the boosted simulations of the parent events separately, as a function of lead time.

Although there are variations between simulations stemming from different parent events, certain patterns are visible. For all panels, the spread between boosted simulations shrinks as the lead time grows shorter, centering around the parent peak. Th

Figure 6. TXx5d of boosted simulations as a function of lead time and parent event. Boosted simulations from parents 1-5 in (a–e) test slice 1 and (f–j) test slice 2. For each parent event a batch of N_{batch} = 100 boosted simulations are generated at every lead time. The median (orange solid line), 5th to 95th percentile range (orange shaded area), and maximum TXx5d (orange plus-sign) are shown. TXx5d for the parent event is represented by a blue horizontal line.

corresponds to boosted simulations being more and more constrained by the dynamics of their parents. Conversely, for longer lead times, the median tends to decrease and the spread tends to grow. In other words, simulations diverge from the parent (and each other) so that the memory of the parent heatwave fades and simulations revert back to the underlying climatology. There are exceptions to this general picture, however, in particular for parent events from Test slice 2. Here, simulations from parent events 1, 2 and 5 seem to show less sensitivity to lead times, with relatively consistent levels of spread and medians and visibly less spread than simulations from other parent events at long lead times.

It is also worth noting that from one lead time to another, both median, spread and maximum values can vary substantially. Additionally, the maximum TXx5d in a given panel is not always from the lead time with the highest median or 95th percentile. This suggests that several lead times might be necessary to better sample the antecedent conditions AC_t.

The effect of the day at which perturbation is performed on the estimation of return periods using the boosting estimator is broken down in more detail in Figure 7. Panel a shows how simulations spread after perturbation: we see that for both test

Figure 7. Link between boosted simulations and perturbation time in (blue) test slice 1 and (orange) test slice 2. (a) Evolution of boosted simulation standard deviation bewteen members with respect to the time elapsed since perturbation. Standard deviation between boosted simulations, divided by the standard deviation of the pre-industrial control simulation: (solid line) median spread across parent events and (shaded area) range of spread across parent events. (b) Frequency of events exceeding T_{ref} in the Boosted Ensemble ($\hat{p}_{T|T_{ref} \mid AC_t}$) as a function of lead time. Error-bars denote the bootstrapped 95% confidence interval.

slice 1 and test slice 2, the spread between simulations initially grows exponentially, before slowing into a linear growth phase and finally saturating to the climatological spread between the simulations in the non-boosted ensemble of the control period. This type of growth is described in ergodic chaos theory, which shows that initially, boosted simulations diverge from their parent exponentially fast with a typical time scale around the inverse of the largest Lyapunov exponent of the system, and has been extensively studied for both climate and weather forecast models (e.g., Trevisan and Palatella, 2011; Vannitsem, 2017). Boosted simulations with short lead times are thus still in the exponential spread stage by the time the heatwave peaks, and will therefore be significantly constrained by the value of the parent event at that time. This means that the dynamics governing the heatwaves in the boosted simulations are structurally too similar to those of the parent event to produce a largely different TXx5d value.

Note that, given the results of Figure 7a, the empirical saturation time after boosting is set to 21 days (see Methods). This means that the maximum Tx5d of the boosted simulations are only assessed in this time range, because afterwards, the memory of the parent, and thus the antecedent conditions leading to a heatwave, are assumed to be lost.

Figure 7b, on the other hand, shows that $\hat{p}_{T|T_{ref} \mid AC_t}$ increases as the lead time grows shorter. Since Figure 3 shows that a higher $\hat{p}_{T|T_{ref} \mid AC_t}$ leads to smaller theoretical errors, shorter lead times should give fewer errors. In other words, a smaller $\hat{p}_{T|T_{ref} \mid AC_t}$ indicates that there are fewer simulations that surpass T_{ref}, and thus less extreme events to find return periods for

Figure 8. Relationship between TXx5d of parent event (blue diamond) and boosted simulations (orange box plot) for (a) test slice 1 and (b) test slice 2. Simulations are pooled together from lead times between 10 and 13 days. Boxplot whiskers are drawn at 1.5 inter-quantile range.

overall. However, as previously mentioned, the longer the lead time, the more each simulation has had time to spread, and thus gain independence from both parents and siblings. Therefore, a trade-off between relative error minimization and sampling more intense events appears. Note that one cannot reliably estimate the spread as a function of lead time, since each boosted simulation may not exactly follow the theoretically predicted ergodic growth speed, in particular due to the extreme nature of the parent events. This is evidenced by the large variation of growth exhibited between boosted simulations from different parents (see Figure 7). Therefore we cannot explicitly state the boundary lead times of this trade-off, in particular since any effort to do so would be both location- and variable-speci c (heatwaves in the PNW region).

The above-described effects do not all set in at the same time across parent events and test slices. Figure 7 shows that while the median spread overlaps well between Test slice 1 and 2 until around 18 days after perturbation, the saturation sets in earlier for Test slice 2. Figure 7 also shows some differences: while both test slices have the same trend, simulations from Test slice 1 seem to have higher $\Phi_{T\ T_{ref}\ jAC_t}$ than for simulations from Test slice 2 until a lead time of around 10 days. Pooling the simulations perturbed at different lead times, like in Figure 5, may therefore provide more robust results. This has the added bene t of adding more independence between boosted simulations.

Finally, Figure 8 shows that there is only a weak relationship between the TXx5d of the parent event and that of its boosted simulations. Indeed, the Spearman correlation between the 90th percentile of the boosted simulations and the 10 parent events, shown in Appendix Figure A5, is only 0.3. Furthermore, Figure 8 shows that the boosted simulations present a remarkable

Figure 9. Return periods of TXx5d in the PNW region estimated with the boosting estimator corrected for non-stationary (climate change) conditions for a lead time range of 18 to 13 days. Estimated return periods for boosted simulations from (a) the entire Parent Ensemble and (b) the entire Parent Ensemble except the most extreme parent (E1) are shown in orange. The TXx5d of the (black dots) 100-member and (blue dots) 30-member Large Ensemble, with a tted GEV law and shown with the median (solid line) and a bootstrapped 95% con dence interval (shaded area). The parent events in the Parent Ensemble are shown with blue diamonds, and the most extreme parent E1, and the most extreme event in the 100-member Large Ensemble are highlighted with stars. The orange horizontal line denotes the reference TXx5d T_{ref}, while the black horizontal line denotes the 2021 ERA5 TXx5d.

variation, not only in the distance from their parent, but also in their median and spread. It is, however, important to note that the boosted simulations generated for this study only belong to a limited number of parent events (N_{parent} = 5 for each test slice), thus preventing us from concluding on a wider basis. Nevertheless, this suggests that a larger Parent Ensemble, i.e. a wider variety of parent events, would not necessarily lead to less extreme boosted simulations, at least not for parents above a certain threshold of intensity.

3.3 Estimating a return period for the Paci c North-West heatwave of 2021

As a further application and a case study, we now use the boosting estimator to estimate a return period, based on the CESM2 climate model, for the record-shattering PNW heatwave of 2021. To do this, we employ a present-climate 30-member Large Ensemble corrected for non-stationarity (see Method Section 2.4) as a reference simulation, from where boosted simulations are generated. The median return period of the 2021 PNW heatwave is calculated using the boosting estimator, and found to be 2500 years with a 95% con dence interval of 2000 to 4000 years. In comparison, the GEV t of the reference 30-member

Large Ensemble and the 100-member Large Ensemble give higher median return period estimates [1 000, 10 000] years and 15 000 [5000, 70 000] years, respectively.

Figure 9a shows all return periods calculated with the boosting estimator for boosted simulations T_{ref} with T_{ref}. Lead times are chosen to be [18, 13], corresponding to the lead times that did not show signi cant constraints by parent events in the pre-industrial control experiment. A lead time by lead time breakdown of return periods is nevertheless shown in Appendix Figure A6. In Figure 9a, we see that the median return period estimates deviate from those of the 30- and 100-member Large Ensembles and their GEV ts for return periods between around 200 and 10 000 years, although this deviation is always within an order of magnitude error of the upper bounds of the GEV t of both the 30- and the 100-member Large Ensemble. This deviation is more present for shorter lead times, not included in Figure 9, but visible in Appendix Figure A6.

One hypothesis that could explain this deviation is that the boosted simulations from the most extreme parent event, hereafter denoted by E1, are biased due to the intensity of the event. Indeed, E1 stands in contrast to the rest of the events in the Parent Ensemble — with a TXx5d of 14.7 C, it is 1.6 C warmer than the second most extreme parent event E2. Remarkably, the 100-member Large Ensemble also presents such an event, with TXx5d of 15.8 C, that is 2.0 C warmer than E2.

To test this hypothesis, Fig. 9b shows return periods calculated excluding boosted simulations from E1. These return period estimates follow the 5th percentile of the uncertainty range of the 100-member Large Ensemble GEV t. Additionally, this effect is unique to the removal of boosted simulations from E1: when removing any other parent event, the results look largely the same (see Appendix Figure A7).

We also see that removing E1 has a strong impact on the GEV distribution t of the 30-member Large Ensemble data: in Figure 9a, the con dence intervals of black and blue ts overlap signi cantly, while in Fig. 9b they are almost disjoint. This highlights the previously stated uncertainty of the bootstrapped con dence interval of the 30-member Large Ensemble, and suggests that the presence of E1 in the Parent Ensemble largely affects the naive return period estimation. Indeed, when removing E1 from the Parent Ensemble, the median return period estimate is in nite, with a con dence interval of [8 000, 110 000] years. Removing E1 when using the boosting estimator, on the other hand, leads to a median estimate of 3500 years with a 95% con dence interval of 2500 to 7000 years, which is within 1000 years of the original estimate.

In other words, had E1 not occurred in the reference simulation, the 2021 PNW heatwave and indeed E1 itself would be judged impossible with only the Large Ensemble data. However, the boosting estimator can provide return periods for extreme events like the 2021 PNW heatwave that depend less on E1, and that, importantly, stay nite. We can thus see that in addition to providing more robust return period estimates for extreme events, Ensemble Boosting also demonstrates that E1 and both the 2021 PNW heatwave events are physically possible, according to the climatology of this model.

To gain a picture of the atmospheric dynamics associated with E1 and the other most intense heatwaves in the 30-member Large Ensemble, the Tx5d anomaly and Z500 contour lines are plotted on the day of each heatwave peak in Figure 10. E1 distinguishes itself from the other top 13 events, and from the 2021 PNW heatwave by the presence of a more distinct cyclonic anomaly in the Z500 eld over the Eastern Paci c, anking the region of interest. Furthermore, the blocking high is centered around the PNW region for both E1 and the 2021 PNW heatwave, while the center of the blocking high shown in the composites is located further South. However, the E1 blocking high is more oblong and rotated due to the presence of the Paci c trough.

Figure 10. Tx5d anomaly [C] and Z500 [m] contour lines on the peak day of each heatwave for (a) the most extreme event in the 30-member Large Ensemble E 1, (b) ERA5 data of the PNW heatwave on 2021-06-28, and composites of E 2 3 and (d) E 2 13, both from the 30-member Large Ensemble. The black box indicates the region of interest. The anomaly calculations are corrected for non-stationarity (see Methods section 2.4.

While the composite maps of E 2 3 and E 2 13 look distinct from that of E 1, individual events within this selection could look different to their mean. Therefore, all 13 events were plotted individually in Appendix Figure A8. While there are events that show a third Z500 trough next to the PNW region (events E 4 and E 11), these are either less distinct from the main troughs forming the block, or fail to create a signi cant blocking high around the PNW region.

Additionally, when analyzing the most extreme from the 100-member Large Ensemble, we see that it presents a similar dynamical situation to that of E 1 (see Appendix Figures A9 and A10) with a third Paci c trough that is not seen in the other top events of the 100-member Large Ensemble.

4 Discussion

4.1 Reduction of errors and computational costs of the boosting estimator compared to the naive estimator

The theoretical derivation of the boosting estimator variance has shown that it can reduce the relative error compared to equivalently expensive naive estimators – and even an estimator based on the full 4000-year control period – given the right con guration and an extreme enough event (see Figure 3 and method section 2.2).