Weather Clim. Dynam., 7, 393–410, 2026
https://doi.org/10.5194/wcd-7-393-2026


Weather and Climate Dynamics
Open Access
EGU

# Constructing extreme heatwave storylines with differentiable climate models

**Tim Whittaker and Alejandro Di Luca**

Centre Étude et simulation du climat à l'échelle régionale (ESCER), Département des Sciences de la Terre et de l'Atmosphère, Université du Québec à Montréal, Montréal, Québec, Canada

**Correspondence:** Tim Whittaker (whittaker.tim@courrier.uqam.ca)

Received: 1 August 2025 – Discussion started: 12 August 2025
Revised: 2 February 2026 – Accepted: 7 February 2026 – Published: 19 February 2026

**Abstract.** Understanding the plausible upper bounds of extreme weather events is essential for risk assessment in a warming climate. Existing methods, based on large ensembles of physics-based models, are often computationally expensive or lack the fidelity needed to simulate rare, high-impact extremes. Here, we present a novel framework that leverages a differentiable hybrid climate model, NeuralGCM, to optimize initial conditions and generate physically consistent worst-case heatwave trajectories. Applied to the 2021 Pacific Northwest heatwave, our method produces heatwave intensity up to 3.7 °C above the most extreme member of a 75-member ensemble. These trajectories feature intensified atmospheric blocking and amplified Rossby wave patterns—hallmarks of severe heat events. Our results demonstrate that differentiable climate models can efficiently explore the upper tails of event likelihoods, providing a powerful new approach for constructing targeted storylines of extreme weather under climate change.

## 1 Introduction

The 2021 Pacific Northwest (PN2021) heatwave shattered historical temperature records, culminating in Lytton, Canada's unprecedented 49.6 °C observation – a 4.6 °C increase over the country's previous record measurement (White et al., 2023; Mass et al., 2024). This event, virtually implausible under preindustrial conditions (Philip et al., 2022), exemplifies a critical challenge in climate science: determining the upper bounds of what is physically possible for different weather extremes under current or future climatic conditions.

The PN2021 heatwave emerged from persistent atmospheric blocking sustained by large-scale Rossby waves that disrupted zonal flow and stalled a high-pressure system over the region (Mass et al., 2024; White et al., 2023). This large-scale setup was fueled by upstream dynamics. Mo et al. (2022) linked it to anomalous atmospheric river activity, while Neal et al. (2022) identified that diabatic heating within the warm conveyor belt of an upstream cyclone provided the necessary Rossby wave activity to establish the block. Once established, the block suppressed cloud formation and drove prolonged subsidence, adiabatically warming near-surface air masses (Loikith and Kalashnikov, 2023). White et al. (2023) corroborated the importance of these mechanisms and estimated via 4 d backward trajectory analysis that diabatic processes accounted for approximately 78 % of the net temperature change of air parcels entering the region, with the remaining $\sim$22 % attributed to adiabatic warming from subsidence. Locally, dry soil conditions further intensified these temperatures through non-linear land-atmosphere interactions (Bartusek et al., 2022; Conrick and Mass, 2023; Schumacher et al., 2022). By studying a 100-member ensemble of PN2021 with varying initial land surface conditions, Duan et al. (2025b) found that variations in antecedent soil moisture led to a spread of approximately 3 °C in peak temperatures, largely driven by regions shifting into a transitional evaporation regime where latent heat flux becomes highly sensitive to soil moisture.

To systematically explore such extremes, storylines are increasingly used, representing physically consistent sequences of weather events that depict how a counterfactual extreme event might occur (Hazeleger et al., 2015; Shepherd, 2019; Sillmann et al., 2021). This approach enables a mecha-

nistic exploration of how minor perturbations can lead to the amplification of extreme events. Here, we use a novel differentiable modeling framework to demonstrate that targeted initial-condition perturbations can further amplify these typical extreme trajectories, giving extreme heatwave storylines.

Identifying storylines for the most extreme weather events is a needle-in-a-haystack problem due to their inherent rarity. The traditional approach is the use of single-model initial-condition large ensembles (Deser et al., 2020; Suarez-Gutierrez et al., 2020; Maher et al., 2021; Diffenbaugh and Davenport, 2021), and more recently, so-called huge ensembles (Mahesh et al., 2024a, b), in which vast numbers of model runs allow the exploration of a wide range of potential outcomes. By systematically increasing ensemble size, the chances of capturing low-probability extremes increase. However, these ensembles are computationally demanding and not very effective at sampling the full range of outcomes. In addition, due to their high computational cost, it is virtually impossible to perform such ensembles using kilometer-scale simulations, which are required to well simulate some types of weather extreme events (e.g., extreme convective precipitation).

In recent years, a number of approaches have been proposed to generate extreme event storylines (Ragone et al., 2018; Plotkin et al., 2019; Webber et al., 2019; Yiou and Jézéquel, 2020; Gessner et al., 2021; Fischer et al., 2023). These methods focus computing resources on specific extreme events, instead of continuous long simulations. Some approaches enhance the likelihood of simulating extreme events by constructing targeted ensembles (Ragone et al., 2018; Webber et al., 2019; Fischer et al., 2023). Fischer et al. (2023) focus on generating an initial condition ensemble of climate model simulations of known extreme events using a method named ensemble boosting. They applied this approach to the PN2021 heatwave and by perturbing the initial conditions using numerical noise for 500 members, they found a 5 d running average of daily maximum temperature anomalies up to 2.9 °C larger than the unperturbed event. Other approaches to construct storylines of extreme events use the large deviation algorithm (Ragone et al., 2018; Ragone and Bouchet, 2021; Noyelle et al., 2025) where an ensemble of simulations is ran and members are periodically pruned or cloned such that an ensemble most likely to lead to an extreme event is generated. Focusing on western European heatwaves and using the large deviation algorithm, Ragone et al. (2018) generated an ensemble which has a mean 2 °C anomaly compared to a control ensemble of 128 members. Other applications of the algorithm showed its ability to identify even more extreme events, with ensembles with mean anomalies of 4 °C (Ragone and Bouchet, 2021). Meanwhile, Plotkin et al. (2019) introduced a variational data assimilation technique, optimized with a 4D-Var inspired method to intensify past extreme tropical cyclones with minimal perturbations. This approach is closely related to the method presented here; however, we leverage automatic differentiation and computationally efficient ML-based models.

Recent advances in machine learning (ML) have led to the development of transformative tools for weather and climate modeling. Neural network architectures like GraphCast (Lam et al., 2023), Pangu-Weather (Bi et al., 2023), FourCastNet (Pathak et al., 2022; Kurth et al., 2023), and FuXi (Chen et al., 2023) have demonstrated forecasting skill comparable to that of traditional numerical weather prediction systems, but at significantly reduced computational costs (Rasp et al., 2024; Pasche et al., 2025; Ennis et al., 2025; Zhang et al., 2025). In addition to their reduced computational costs, these models by construction allow us to define optimization problems on them that can be solved through gradient-based optimizers. Rasp et al. (2024) introduced a standardized benchmark to compare the various ML models against ERA5 and the European Centre for Medium-Range Weather Forecasts's (ECMWF) integrated forecast system (IFS). Using this benchmark, it is shown that deterministic, data-driven methods such as Pangu-Weather, GraphCast, and FuXi result in similar root-mean-square error (RMSE) in forecasting near-surface temperature, wind, and pressure up to 10 d ahead. However, their forecast skill deteriorates rapidly for longer lead times, resulting in overly smoothed predictions. Using three case studies, Pasche et al. (2025) evaluated GraphCast, Pangu-Weather and FourCastNet against ERA5 reanalysis and ECMWF's IFS for the PN2021 heatwave, the 2023 South Asian humid heatwave and a 2021 North American winter storm. They find that all data-driven models systematically underestimate the peak 2m temperature during the PN2021 heatwave, with root mean square error (RMSE) values at grid points near Vancouver, Seattle, and Portland exceeding twice the 10 d IFS error and reaching up to four times that value in Portland. During the South Asian humid heatwave, data-driven forecasts of heat index computed from 2 m air temperature and 1000 hPa relative humidity underpredicted observed peaks more strongly than IFS, particularly over Bangladesh. For the North American winter storm, data-driven forecasts of wind chill at College Station, Texas, achieved lower peak errors than IFS, with Pangu-Weather and GraphCast outperforming the operational model.

An alternative to purely data-driven approaches is the use of hybrid models, such as NeuralGCM (Kochkov et al., 2024), which combines a traditional dynamical core with ML components. (Duan et al., 2025a) have shown NeuralGCM's ability to hindcast the PN2021 heatwave, though due to the lack of processes (such as land-atmosphere feedbacks), the intensity of the heatwaves tends to be underestimated. Similarly to the purely data-driven models, NeuralGCM produces surface variables forecasts with skill comparable to that of the ECMWF IFS system Rasp et al. (2024). Moreover, the use of the dynamical core both prevents the evolved fields from being overly smoothed and enhances numerical stability. These benefits allow for longer time integrations and

make the model suitable for climate studies (Kochkov et al., 2024).

These new types of models are by construction differentiable through automatic differentiation (Gelbrecht et al., 2023). The automatic differentiation property enables efficient optimization, allowing gradient-based exploration storylines in high-dimensional climate models. This is in line with many new extreme event opportunities enabled by ML models (Materia et al., 2024; Camps-Valls et al., 2025). Leveraging automatic differentiation, recent studies have implemented variational data assimilation techniques using neural networks, with applications ranging from toy models, such as the Lorenz 96 system (Lorenz, 2006), to reduced-order physical representations of the atmosphere (Solvik et al., 2025; Manshausen et al., 2024). Additionally, Vonich and Hakim (2024) demonstrated that the differentiability of GraphCast allows for a more accurate reconstruction of the initial conditions that led to the PN2021 heatwave compared to using ERA5 reanalysis data. Baño-Medina et al. (2025) explores the use of ML models and automatic differentiation to perform sensitivity analysis of the initial conditions leading to the development of cyclone Xynthia. Their findings suggest that gradients computed from the data-driven weather model at a 36-hour lead time exhibit sensitivity structures that closely resemble those generated by the adjoint of a dynamical model. In other words, the evolved perturbations from both approaches lead to similar impacts on the cyclone's evolution.

In this study, we focus on the PN2021 heatwave event due to its well-documented synoptic drivers, and its prevalence in extreme event studies (Lucarini et al., 2023; Fischer et al., 2023; White et al., 2023; Philip et al., 2022). We use the automatic differentiation feature of the NeuralGCM model to optimize perturbed initial conditions, and we identify trajectories where enhanced geopotential height anomalies intensify downstream near-surface temperature extremes. These storylines reveal heatwave intensity increases of 3.7 °C beyond the extreme temperatures obtained from a 75-member ensemble run using NeuralGCM for the event, analogous to the ensemble boosting approach (Fischer et al., 2023). Our results demonstrate the potential of differentiable hybrid models for investigating worst-case scenarios, offering a computationally efficient alternative to traditional, computationally expensive, large ensembles.

This paper is organized as follows. Section 2 introduces an optimization problem whose minimization yields extreme heatwaves and describes how NeuralGCM is used to solve it. Section 3 presents the optimized heatwave storylines in comparison with an ensemble run. Section 4 discusses the implication of the method and future directions. Finally, Sect. 5 concludes the paper.

## 2 Methods

### 2.1 Initial Conditions Optimization Problem

Our goal is to find the worst-case physically plausible heatwave trajectory our model can produce. To achieve this, we must find the specific, small perturbations to a known initial state that will evolve into the most extreme event. This search is formulated as an optimization problem, where we define a loss function that the model will automatically minimize in an iterative way to find these optimal initial-state perturbations. Formally, a suitable loss function for our problem is one that

1. maximizes a target extreme event, and
2. minimizes the introduced perturbation.

The optimization process is framed through a continuous-time dynamical system. For conceptual clarity, we describe the problem using an ordinary differential equation system:

$$\dot{\boldsymbol{x}} = b\left(\boldsymbol{x}(t, \boldsymbol{x}_0)\right), \tag{1}$$

with $b$ representing some nonlinear operator, $t$ the time and $\boldsymbol{x}_0$ some initial state $t_0$. The solution is given by $\boldsymbol{X}(t) = S_t \boldsymbol{x}_0$ where $S_t$ is the evolution operator up to time $t$. The core aim of the optimization problem is to identify the initial conditions $\boldsymbol{x}_0$ that drive the dynamical system toward an extreme desired state, represented by a target observable $\mathcal{O}(\boldsymbol{X}(t))$. Given a baseline initial state $\boldsymbol{x}_0^b$, we define a perturbation $\Delta\boldsymbol{x}_0 = \boldsymbol{x}_0^b - \boldsymbol{x}_0$. The optimization problem is formulated in terms of minimizing the following loss function:

$$\mathcal{L}\left(\boldsymbol{X}(t), \Delta\boldsymbol{x}_0\right) = F\left(\mathcal{O}(\boldsymbol{X}(t))\right) + \boldsymbol{\lambda} \cdot \Delta\boldsymbol{x}_0^2, \tag{2}$$

where the first term is designed to favor more extreme values of the observable by applying a cost function $F$ to the outcome $\mathcal{O}(\boldsymbol{X}(t))$, and the second term, scaled by the regularization parameter $\boldsymbol{\lambda}$, penalizes the magnitude of the initial perturbation $\Delta\boldsymbol{x}_0^i$. This formulation balances the competing objectives of inducing a rare event and keeping the initial perturbation sufficiently small.

In particular, we pick our observable $(\mathcal{O}(\boldsymbol{X}(t)))$ to be the temperature over a domain $\mathcal{D}$ and over a period of time $\tau$ at the 1000 hPa pressure level of the model $(\int_0^\tau \int_D T_{1000}(\phi, \theta, t)\mathrm{d}\phi\,\mathrm{d}\theta\,\mathrm{d}t)$. Multiple functions $F(\mathcal{O}(\boldsymbol{X})$ can be considered, but our main results use $F(X) = \frac{c}{X}$ which gives us the loss:

$$\mathcal{L}(T_{1000}, \Delta x_0) = \underbrace{\beta \frac{T_{\mathrm{ref}}}{\frac{1}{\tau|D|}\int_0^\tau \int_D T_{1000}(\phi, \theta, t)\,\mathrm{d}\phi\,\mathrm{d}\theta\,\mathrm{d}t}}_{\text{Temperature objective term}} + \underbrace{\sum_i \lambda_i \frac{(\Delta x_{0,i})^2}{(\Delta x_{\mathrm{ref},i})^2}}_{\text{Perturbation penalty term}} \tag{3}$$

where $\mathcal{D}$ corresponds to the region shown in Fig. 4, and $\tau$ is set to 5 d. This 5 d period was chosen to fully encompass the 3 peak days of the PN2021 event, with a 2 d buffer at the end, which we found aided in optimization. The terms in the loss function are normalized by their initial means, with $T_{\text{ref}}$ representing a characteristic temperature scale and $\Delta x_{\text{ref},i}$ denoting a reference perturbation scale for each perturbed variable $i = \{$Temperature, Surface Pressure, Vorticity, Divergence, Specific Humidity, Specific Cloud Ice Water Content, Specific Cloud Liquid Water Content.$\}$. The normalization scale for each perturbed variable, $\Delta x_{\text{ref},i}$, is defined as the absolute mean of each respective initial field. This ensures that each term is of similar magnitude.

Once the simulations are optimized, we evaluate their success through an intensity metric for the heatwaves. We define a heatwave event as a period during which the daily temperature exceeds the 99th percentile threshold for consecutive days (Comeau et al., 2025). This definition relies on the persistence of temperature extremes (see also heatwave intensity definition); if the temperature drops below the threshold for even a single day, the event is considered terminated, and any subsequent exceedances are treated as distinct, separate events. The intensity of the heatwave is measured by the average exceedance of the temperature above the threshold over the duration of the event. Specifically, if $L$ denotes the length of the event, $T_i$ the mean temperature time series over a region, and $T_{\text{thresh}}$ the 99th percentile threshold, then the intensity $I$ is defined as

$$I = \frac{1}{L}\sum_{i=1}^{L}(T_i - T_{\text{thresh}})\,. \tag{4}$$

In our analysis, we compare the intensity, $I$, of the heatwaves from the optimized runs with those from the ensemble runs over the targeted 5 d of the optimization process.

### 2.2 Numerical Implementation using NeuralGCM

To simulate the dynamics and evaluate the loss function, we use the NeuralGCM model (Kochkov et al., 2024). Most of the experiments are performed with a horizontal grid spacing of 2.8° (denoted as NeuralGCM2.8) because it is more computationally tractable and because a 40-year climate simulation is readily available at this resolution. For sensitivity analysis, we also consider a horizontal grid spacing of 1.4° (denoted as NeuralGCM1.4). NeuralGCM employs a dynamical core to solve the primitive equations using a semi-implicit time-integration scheme and a spectral method. Physical processes on the other hand are emulated by learned physics through a neural network.

NeuralGCM has been implemented in JAX (Bradbury et al., 2018) and supports automatic differentiation. This enables the computation of gradients with respect to both initial conditions and internal system parameters, facilitating backpropagation through the physical dynamics and neural network components. In this work, we compute gradients only with respect to the initial variables involved in the dynamical core of NeuralGCM, keeping all other parameters fixed. The loss function, as defined in Eq. (3), is minimized using gradient descent, specifically with the Adam optimizer from Optax (DeepMind et al., 2020). The optimal perturbations are applied to the spherical harmonic coefficients representation of the variables. We choose NeuralGCM over other possible models because it has demonstrated competitive forecast skill for temperatures up to 10 d lead times, contains a dynamical core, and relies on a single initial condition.

Although the model runs efficiently on a single GPU with relatively low memory requirements, gradient computation demands substantial memory, scaling rapidly with the number of time steps. To address this, we employ gradient checkpointing and chunking strategies to manage memory usage. These techniques store only essential intermediate values during the forward pass, recomputing them during the backward pass to reduce memory overhead (Kochkov et al., 2024). The optimization scheme on the 2.8° model runs on a 16 GB A4000 NVIDIA GPU, whereas the 1.4° model necessitates a 40 GB A100 NVIDIA GPU.

We investigate extreme events by perturbing the initial conditions primarily around the PN2021 event using data from the ERA5 reanalysis (Hersbach et al., 2020). We conducted two independent optimization experiments, hereafter referred to as "EXP50" and "EXP75". Their configurations – including the learning rate ($\alpha$), loss-function weights ($\beta$, $\lambda_i$), forecast lead times, initialization dates, and number of gradient descent steps ($N$) – are detailed in Table 1. These parameters were selected via an experimental approach analogous to machine learning hyperparameter tuning, as an exhaustive automated search would be computationally prohibitive. We initially selected $N = 75$ to establish a baseline comparable in computational cost to a 75-member ensemble. Subsequently, we performed the $N = 50$ experiment to assess whether similar results could be achieved with fewer resources. This required retuning the $\lambda_i$ parameters; generally, a larger $N$ implies a longer search time, allowing perturbations to grow larger, which in turn necessitates a higher $\lambda$ to constrain their size. Finally, forecast lead times were chosen to strike a balance: sufficiently close to the event to ensure forecastability, yet distant enough to allow the introduced perturbations adequate time to evolve.

The optimized simulations are compared to an ensemble run of the event using the stochastic version of NeuralGCM. This ensemble consists of 75 members. Unlike our approach, which perturbs the initial conditions (inputs to the model), the stochastic model introduces perturbations within the learned physics module. As a result, the perturbations are effectively introduced one time step apart. Additionally, our method perturbs surface pressure, which is not perturbed in the stochastic model. More details about the stochastic model can be found in Kochkov et al. (2024).

**Table 1.** Parameters used during the optimization process. Each row corresponds to one experiment. The coefficients $\lambda_T$, $\lambda_{SP}$, $\lambda_\delta$, $\lambda_\zeta$, $\lambda_{SH}$, $\lambda_{SCIWC}$, and $\lambda_{SCLWC}$ control the relative weight of the temperature term, the surface pressure term, the divergence term, the vorticity term, the specific humidity term, and the ice and liquid cloud water terms in the loss function. The parameter $\beta$ sets the strength of the temperature objective term. The number of iteration steps $N$ differs between the two experiments in order to explore the effect of longer and shorter optimization procedures while all other settings are kept fixed. The quantity $\tau$ denotes the forecast lead time used when computing the loss.

| Experiment name | $\alpha$ | $\beta$ | $\lambda_T$ | $\lambda_{SP}$ | $\lambda_\delta$ | $\lambda_\zeta$ | $\lambda_{SH}$ | $\lambda_{SCIWC}$ | $\lambda_{SCLWC}$ | Initial Date | $\tau$ | Total integration time | $N$ |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| EXP50 | $10^{-9}$ | 20 | 200 | 20 | 2000 | 2000 | 200 | 20 | 20 | 21 June 2021 | 5 d | 11 d | 50 |
| EXP75 | $10^{-9}$ | 10 | 100 | 10 | 1000 | 1000 | 100 | 10 | 10 | 21 June 2021 | 5 d | 11 d | 75 |

### 2.3 Initial Condition Perturbations

Table 2 presents the maximum perturbations applied to the initial conditions, alongside the range of values sampled in the 75-member ensemble simulation. The range is computed by finding the maximum perturbation of all the ensemble members with respect to the ensemble mean. Overall, the applied perturbations during optimization remain within or below the range represented in the stochastic ensemble. Although direct visualization of the perturbations is challenging due to their high dimensionality, their spatial spectra provide useful insight and are presented in Sect. 3.5.

## 3 Results

### 3.1 NeuralGCM temperature evaluation

We first evaluate the ability of the NeuralGCM2.8 model to simulate summer (June, July, August) temperatures compared to ERA5. Figure 1a) presents the 6-hourly temperature distribution for a NeuralGCM2.8 simulation and the ERA5 data over the 1981–2020 40-year period, averaged within the domain of interest (highlighted in the blue box in Fig. 4). The NeuralGCM2.8 simulation closely approximates the ERA5 distribution, demonstrating its ability to reproduce key statistical characteristics of the temperature distribution in this region. We highlight the 95th and 99th percentile values for both the model and ERA5. We note that the model has slightly colder hot extremes than ERA5.

Next, the ability of the NeuralGCM2.8 and NeuralGCM1.4 model to forecast the PN2021 heatwave against the ERA5 reanalysis data is evaluated. Figure 1b) shows NeuralGCM2.8 forecasted surface temperatures during the PN2021 heatwave for five lead times ranging from 10 to 2 d. At a 10 d lead time, the NeuralGCM2.8 predictions follow closely the ERA5 data until about day 8, where they deviate leading to a lack of heatwave and extreme temperatures. At an 8 d lead time, the simulation substantially enhances temperature during the heatwave period but still shows large underestimations of peaks intensities, with differences of about 6 °C. At 2, 4, and 6 d lead times, the NeuralGCM2.8 model captures well the general pattern of temperature variations shown by ERA5, including the occurrence of very high temperatures during the heatwave event. However, most forecasts underestimate the peak magnitude compared to ERA5 by a few degrees Celsius, particularly during the days after the peak of the event.

This underestimation of the extreme heat is to our knowledge, two folds: (1) there seems to be a dependence on capturing the extreme with the coarseness of the model, when we increase the resolution to the 1.4° model, the prediction quality improves (see Fig. 1c) and (2) other studies have evaluated the ability of simulating extreme heatwave storylines and found that the model lacking processes, such land-surface feedbacks led to under representation of extreme (Duan et al., 2025a).

### 3.2 Optimizing extreme temperatures

We optimize the initial conditions of the NeuralGCM2.8 model starting from 21 June 2021 (corresponding to a lead time of 8 d to the peak of PN2021; see Sect. 2 for details) and run the simulation forward for 11 d. An 8 d lead time strikes a balance between two requirements: keeping the event within the model's predictable window and allowing enough time for small perturbations to develop. The optimization is performed using gradient descent over 75 steps to solve Eq. (3), targeting the last 5 d of the event (see the gray shaded area in Fig. 1). The full set of parameters used in the optimization is provided in Table 1. Figure 2 shows the differences in 500 hPa geopotential height ($Z_{500}$; top row) and the 1000 hPa temperature ($T_{1000}$; bottom row) between the optimized (OPT) and control (CTL) trajectories. In addition, the $Z_{500}$ and $T_{1000}$ of the optimized simulation are shown in dark contour lines. The early day conditions (T-6 days) show minimal differences, with anomalies amplifying progressively as we get closer to the peak day. Positive and negative differences in $Z_{500}$ are generally observed in association with ridges and troughs, respectively, indicating that the optimized simulation amplifies the hemispheric wave amplitudes. Specifically, over the targeted region, there is a clear increase in $Z_{500}$ and $T_{1000}$, with the largest increases centered over the targeted region.

**Table 2.** Maximum perturbations over the full 3D fields for different run sizes compared to the range of perturbations applied on the ensemble run by the stochastic model.

| Quantity/Experiment | EXP50 | EXP75 | 75-member ensemble |
|---|---|---|---|
| No. of steps | 50 | 75 | – |
| Surface Pressure (Pa) | 0.69 | 0.47 | 0.0 |
| Specific Humidity | $2.34 \times 10^{-3}$ | $1.11 \times 10^{-3}$ | $3.20 \times 10^{-3}$ |
| Specific Cloud Ice Water Content ($kg\,kg^{-1}$) | $7.61 \times 10^{-6}$ | $5.19 \times 10^{-6}$ | $4.35 \times 10^{-5}$ |
| Specific Cloud Liquid Water Content ($kg\,kg^{-1}$) | $1.50 \times 10^{-5}$ | $2.42 \times 10^{-5}$ | $7.61 \times 10^{-5}$ |
| Temperature (K) | 4.83 | 4.99 | 7.60 |
| $U$ component of windspeed ($m\,s^{-1}$) | 8.37 | 5.59 | 12.70 |
| $V$ component of windspeed ($m\,s^{-1}$) | 4.77 | 4.12 | 7.94 |

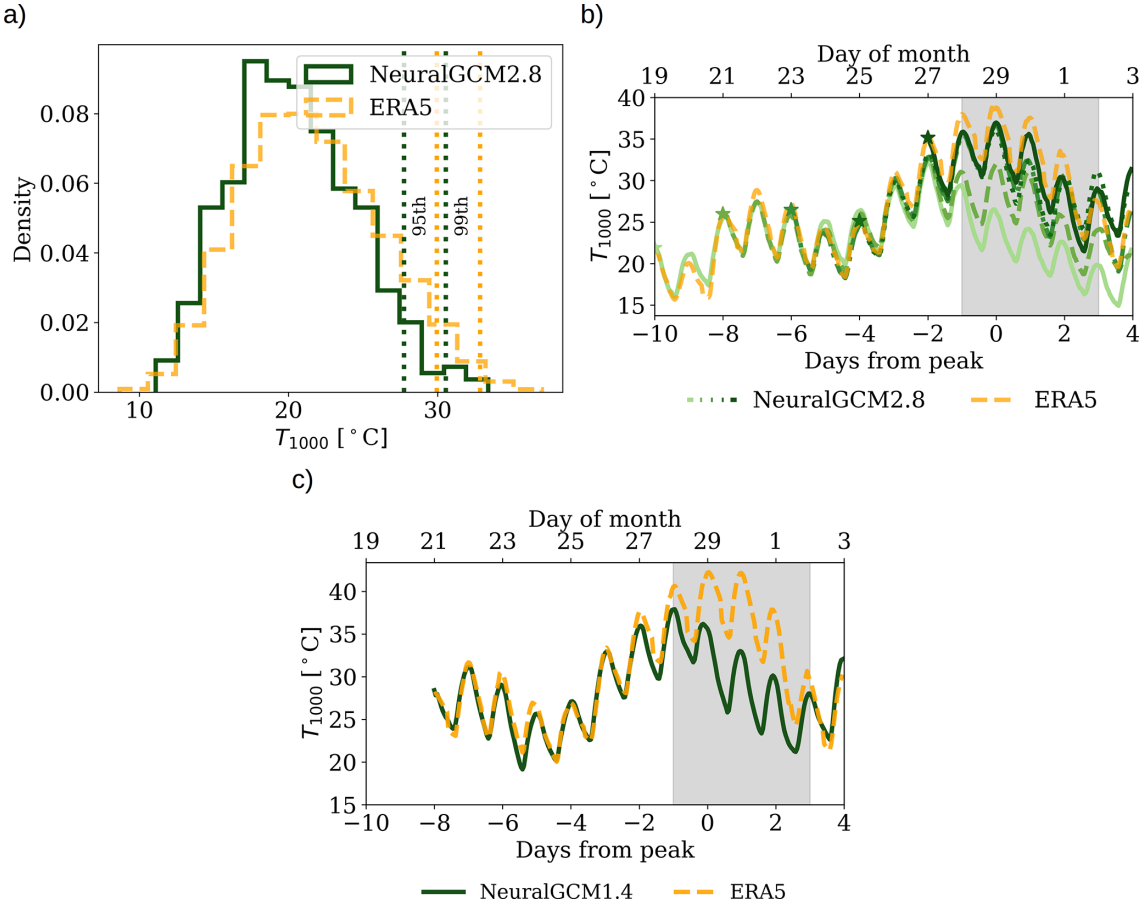


**Figure 1.** **(a)** Histograms of 6-hourly JJA temperature values from a 40-year NeuralGCM2.8 simulation (green) and from ERA5 (orange) reanalysis over the study domain outlined in Fig. 4. Dashed line indicates the 95th and 99th percentiles. **(b)** Time series of temperature forecasts at 1000 hPa from NeuralGCM2.8 with 10, 8, 6, 4, and 2 d lead times (green colored lines) compared with ERA5 reanalysis data (orange line) for the PN2021 heatwave. Grey area highlights the targeted time range for the optimization process. **(c)** Time series of temperature forecasts at 1000 hPa from NeuralGCM1.4 with 8 d lead times.

We examine the 500 hPa geopotential height along a fixed latitude (latitude = 57.2°) in Fig. 3. The wave patterns produced by the optimized simulation are compared to those from the control simulation, along with their respective spectral characteristics during the last 3 d of the event. Notably, the geopotential height near the heatwave region is significantly higher in the optimized simulation compared to the control one. Both the control and optimized simulations show signs of a persistent wavenumber three wave. In the spectral amplitude, the largest differences in spectral amplitude occur for wavenumbers 2-5, which are typically associated with heatwave events. Specifically, the largest dif-

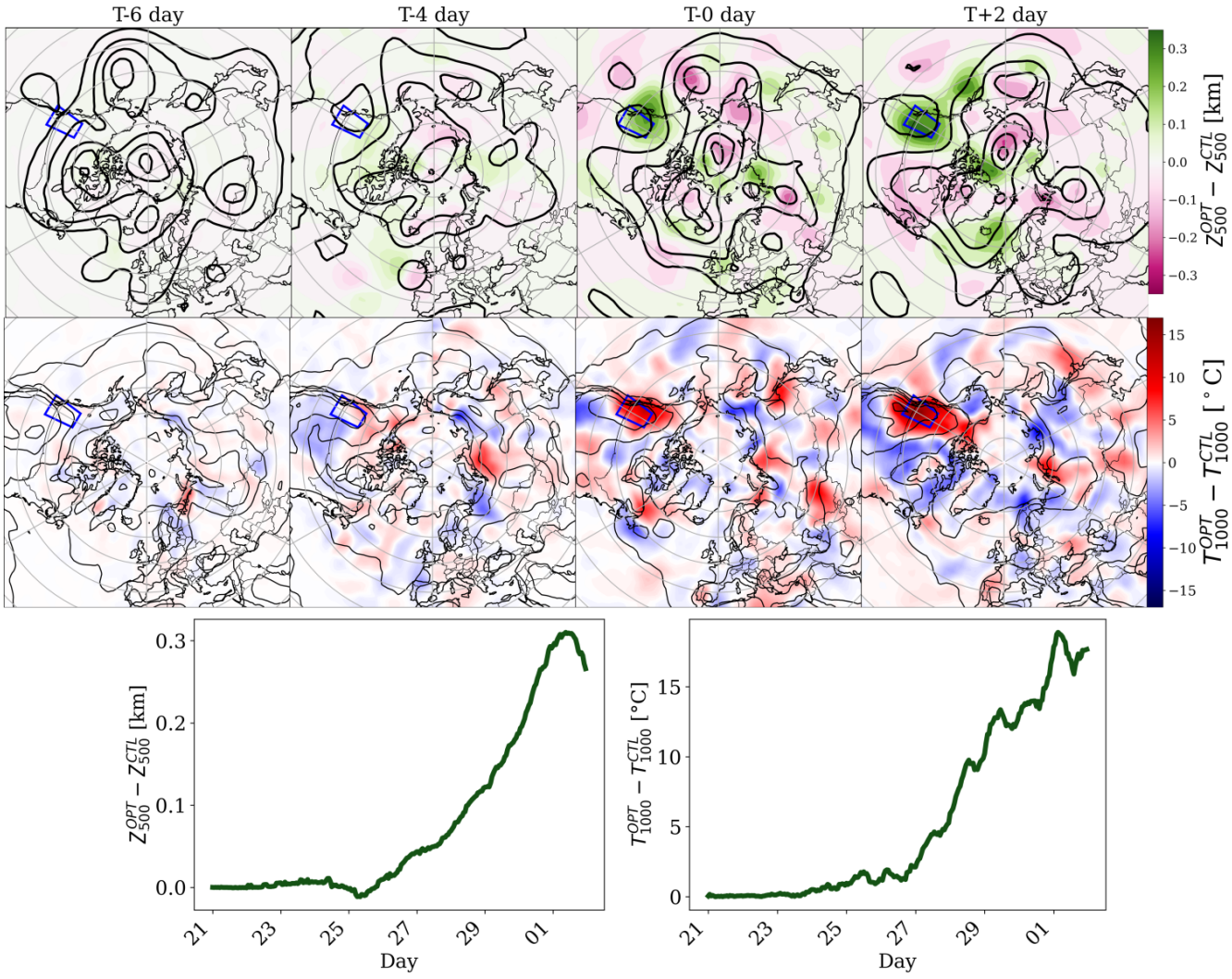


**Figure 2.** Top row: evolution of the difference in 500 hPa geopotential height ($\Delta Z_{500}$, in km) between the optimized simulation and the control run for EXP75. Black contours (optimized run) outline the amplified Rossby wave pattern, with deeper troughs and higher ridges compared to the control. Middle row: The difference in 1000 hPa temperature ($\Delta T_{1000}$, in °C) between the optimized and control run. Bottom row: the difference between the 500 hPa geopotential height and 1000 hPa temperature averaged over the target domain.

ferences are observed at wavenumber 3, followed by 2 and 4, where the optimized simulation exhibits greater power than the control simulation. While some differences are also present at higher wavenumbers, their magnitude is substantially smaller.

The optimization process relies on gradient descent (see Sect. 2 for details), which requires choosing the number of gradient descent steps. Figure 4 shows hourly time series of $T_{1000}$ (Fig. 4a) and $Z_{500}$ (Fig. 4c) for two optimized trajectories with $N = 50$ (EXP50; solid green line) and $N = 75$ (EXP75; dashed green line) steps, alongside a 75-member ensemble (grey lines) and its mean (black thick line), all initialized 8 d before the peak of the event. Notably, the $T_{1000}$ time series (Fig. 4a) reveals that both optimized trajectories attain values beyond the range exhibited by any individual ensemble member. In other words, the proposed method allows us to find extreme temperature values that are more extreme than those found using a 75-member ensemble using only 50 iterations (a 33 % reduction in computational cost relative to generating the 75-member ensemble, calculated as $(75 - 50)/75$). Notably, this more efficient 50-step optimization run produces a trajectory more extreme than any member of the 75-member ensemble. Specifically, the trajectory from 50 steps reaches a peak temperature of 37.0 °C, while the trajectory after 75 steps attains 38.9 °C. Compared to the mean of the ensemble, we reach anomalies of 14.0 °C. For the 500 hPa geopotential height (Fig. 4c), both optimized trajectories show similarly elevated values, once again exceeding the range spanned by the 75-member ensemble. Importantly, the trajectory from the 75-step optimized run main-

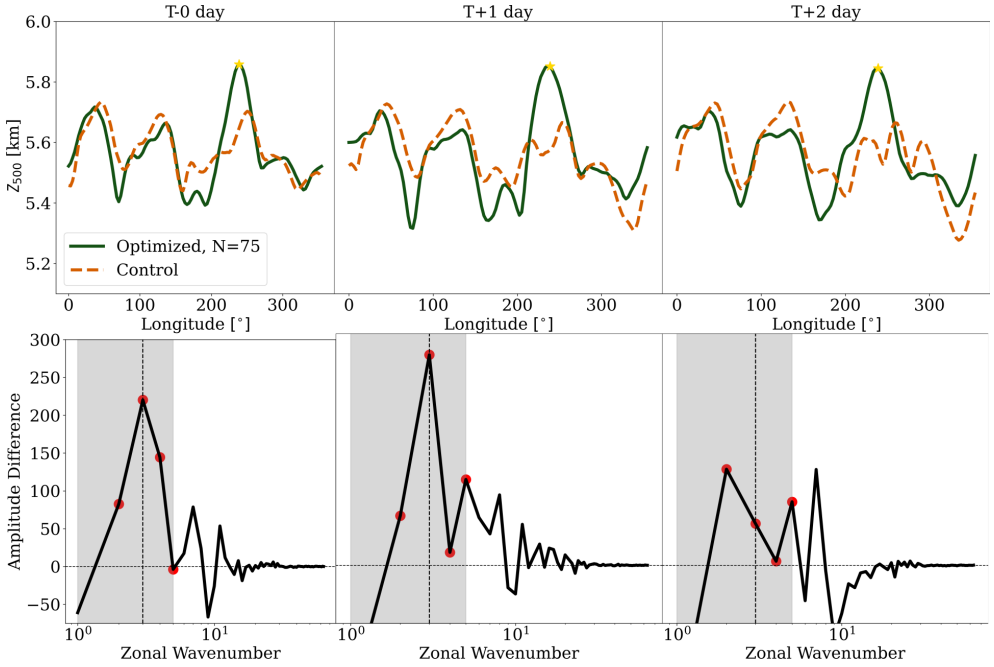


**Figure 3.** Top row: cross section of 500 hPa geopotential along a fixed latitude for EXP75. Bottom row: the amplitude difference of the Fourier spectrum, including wavenumbers 1–5 highlighted in gray. Red dots highlight wavenumbers 2–5.

tains a more sustained increase of the 500 hPa geopotential height compared to that from the 50-step run. The measured intensity (Fig. 4e) and length of the event (Fig. 4f) are increased in both optimized runs compared to the ensemble mean. Both optimized runs produce a 6 d-long heatwave event, differing only in intensity, with the 75-step run having a 1.0°C higher intensity than the 50-step run. Notably, both optimized solutions exceed the intensities spanned within the 75-members ensemble. Figure 4b, d present the temperature and geopotential fields from the 75-iteration run. The temperature pattern features a maximum over the targeted region, albeit slightly to the south, while the 500 hPa geopotential height field exhibits an anticyclone directly overhead.

As shown in Fig. 5, both optimized solution yields fields that are consistent with the ERA5 depiction of the PN2021 event. The ERA5 data shows a maximum temperature of nearly 40 °C on 29 June, similar to the temperature attained by the optimized trajectory of nearly 40 °C for EXP75. The geopotential height of the optimized trajectory, on the other hand, reaches a higher value than ERA5, exceeding 5.9 km by 1 July for EXP75 compared to ERA5's peak of approximately 5.85 km on 29 June which then declines. While ERA5 shows a clear decline in both temperature and geopotential height after the 29 June peak, the optimized trajectories maintain elevated values through 1 July. This sustained behavior is a direct consequence of the loss function, which maximizes the integrated temperature over the 5 d target window ($\tau = 5$ d) rather than targeting a single peak day. Consequently, the optimizer identifies initial conditions that not only amplify the intensity of the event but also extend the duration of the blocking pattern that maintains it. This contrasts with the natural evolution seen in ERA5, where the ridge weakens and temperatures decline within 1–2 d after the peak.

### 3.3 Sensitivity of other variables

Figure 6 shows optimized trajectories for near-surface winds (zonal, $U$ and meridional, $V$, components), specific humidity, and surface pressure. In the optimized run, the zonal wind consistently lies at the lower end of the ensemble spread during the event, and specific humidity likewise tracks near the lower end. Such concurrent reductions in near-surface wind speeds and humidity are consistent with the physical mechanisms that underlie heatwave intensification. In contrast, the meridional wind and surface pressure exhibit only a slight positive anomaly above the ensemble mean. All optimized trajectories remain entirely within the bounds defined by the non-optimized ensemble members. While the variables are within the range of the ensemble envelope (i.e., they are not extreme), there might be a confluence of factors that lead to the extreme. For comparison, the equivalent figure using the ERA5 data for the PN2021 event is presented in Fig. 7. The specific humidity, and winds are within the envelop of the NeuralGCM2.8 ensemble. The surface pressure, on the other hand, has a positive bias in NeuralGCM likely due to the representation of the surface in the coarse NeuralGCM2.8 model.

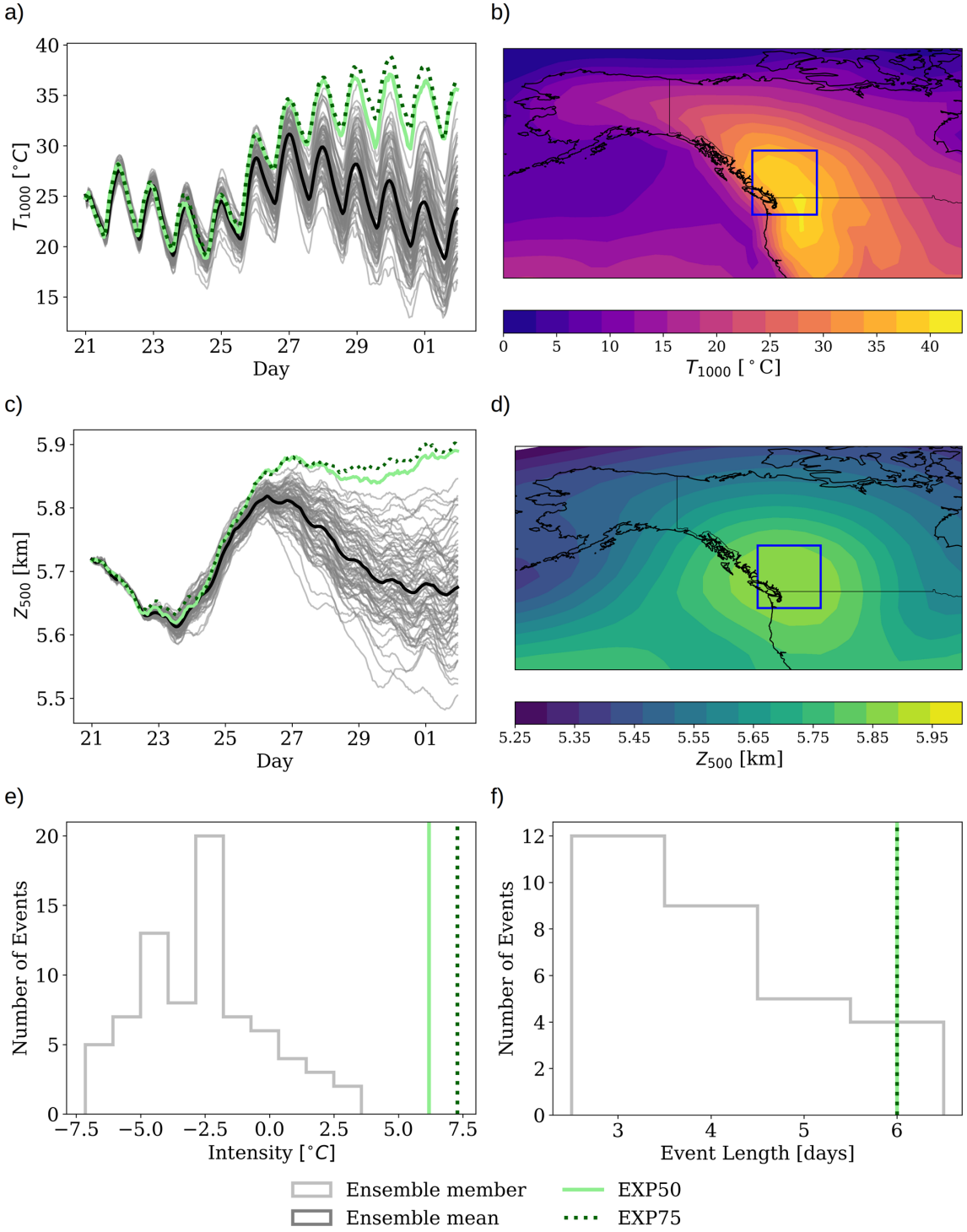


**Figure 4.** **(a)** Time series of 1000 hPa temperature for two optimized trajectories (50 and 75 steps) and a 75-member ensemble with its mean. **(b)** Spatial map of average temperature anomalies from the 75-step run. **(c)** Time series of 500 hPa geopotential height for the same set of simulations. **(d)** Spatial map of 500 hPa geopotential height anomalies during the event period. **(e, f)** Time series of heatwave intensity (defined by Eq. 4) and duration for the ensemble and optimized cases.

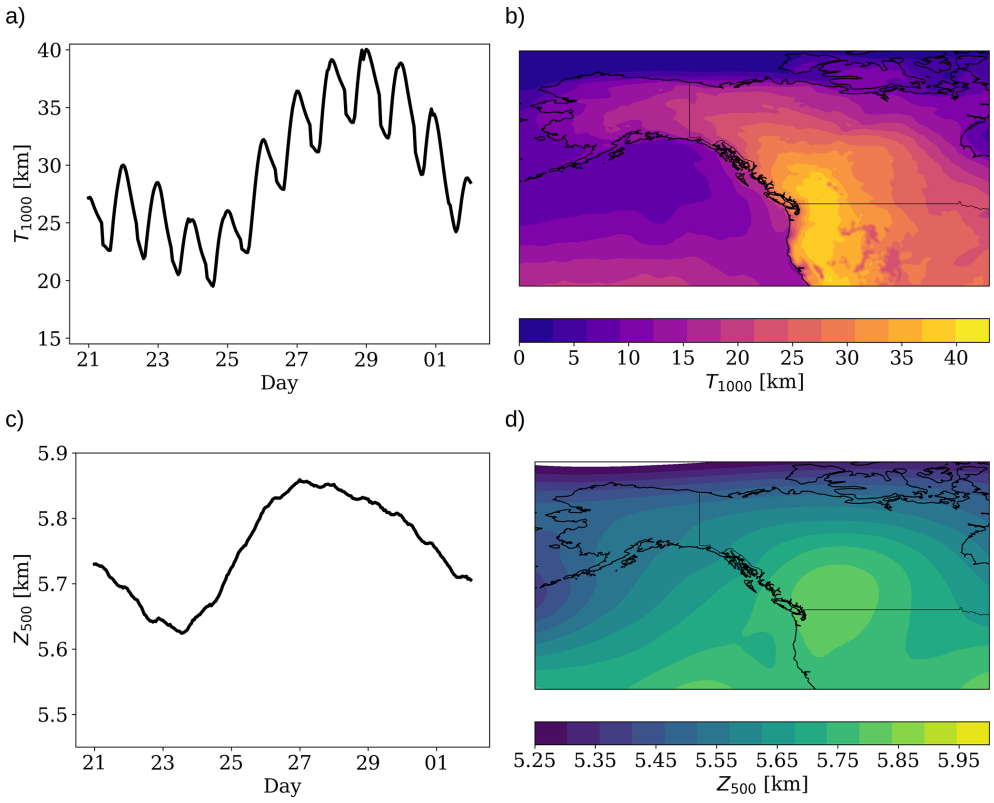


**Figure 5.** Same as panels **(a)**–**(d)** in Fig. 4 but using the ERA5 reanalysis data. **(a)** Time series of 1000 hPa temperature for ERA5 for the PN2021. **(b)** Spatial map of average temperature anomalies from the ERA5 data. **(c)** Time series of 500 hPa geopotential height for ERA5. **(d)** Spatial map of 500 hPa geopotential height anomalies during the event period.

### 3.4 Sensitivity to NeuralGCM resolution

To test how resolution affects our optimization, we reran the optimization problem on NeuralGCM at a finer 1.4° resolution (Fig. 8). The same parameters as in Table 1 are used for this set of experiments. In the unperturbed control, the high-resolution ensemble mean reduces the warm bias against ERA5 and more accurately captures the peak and decay of the PN2021 heatwave. Once optimized, the 1.4° run again produces peak surface-temperature anomalies that exceed the 75-member ensemble maximum, and 500 hPa geopotential-height anomalies that surpass the control even more than the 2.8° case. These enhanced temperature and geopotential height anomalies persist through the extended target period, demonstrating that the optimization delivers sustained extremes even at higher resolution.

### 3.5 Initial condition perturbations spectra

The perturbations introduced in the optimized runs are fully three-dimensional and span all horizontal and vertical levels. Due to this high dimensionality, it is challenging to visualize their full structure directly. To provide some insight into their characteristics, we show in Fig. 9 the spatial spectrum of the perturbations at selected vertical levels (1000, 850, 500 and 200 hPa) for four variables: geopotential height, kinetic energy, specific humidity, and temperature. This representation highlights the dominant spatial scales of the perturbations across the domain. The perturbations seem to have some spatial structure. In particular, geopotential, kinetic energy, and temperature exhibit comparable energy magnitudes across all model levels, with large-scale perturbations dominating the spectrum. Specific humidity, conversely, displays significantly lower energy levels at lower altitudes, though it remains dominated by low-wavenumber perturbations. This structured distribution stands in contrast to a white-noise scenario where the resulting spectrum would be flat across all scales. A detailed analysis of their full spatial structure is beyond the scope of this work. The maps of the associated spectrums are available in the Supplement.

## 4 Discussion

Our findings demonstrate that differentiable climate models, exemplified by NeuralGCM, offer a powerful tool for constructing extreme heatwave storylines through gradient-based optimization of initial conditions. By perturbing initial conditions, we identified alternative trajectories with slightly different synoptic-scale conditions that amplify the PN2021

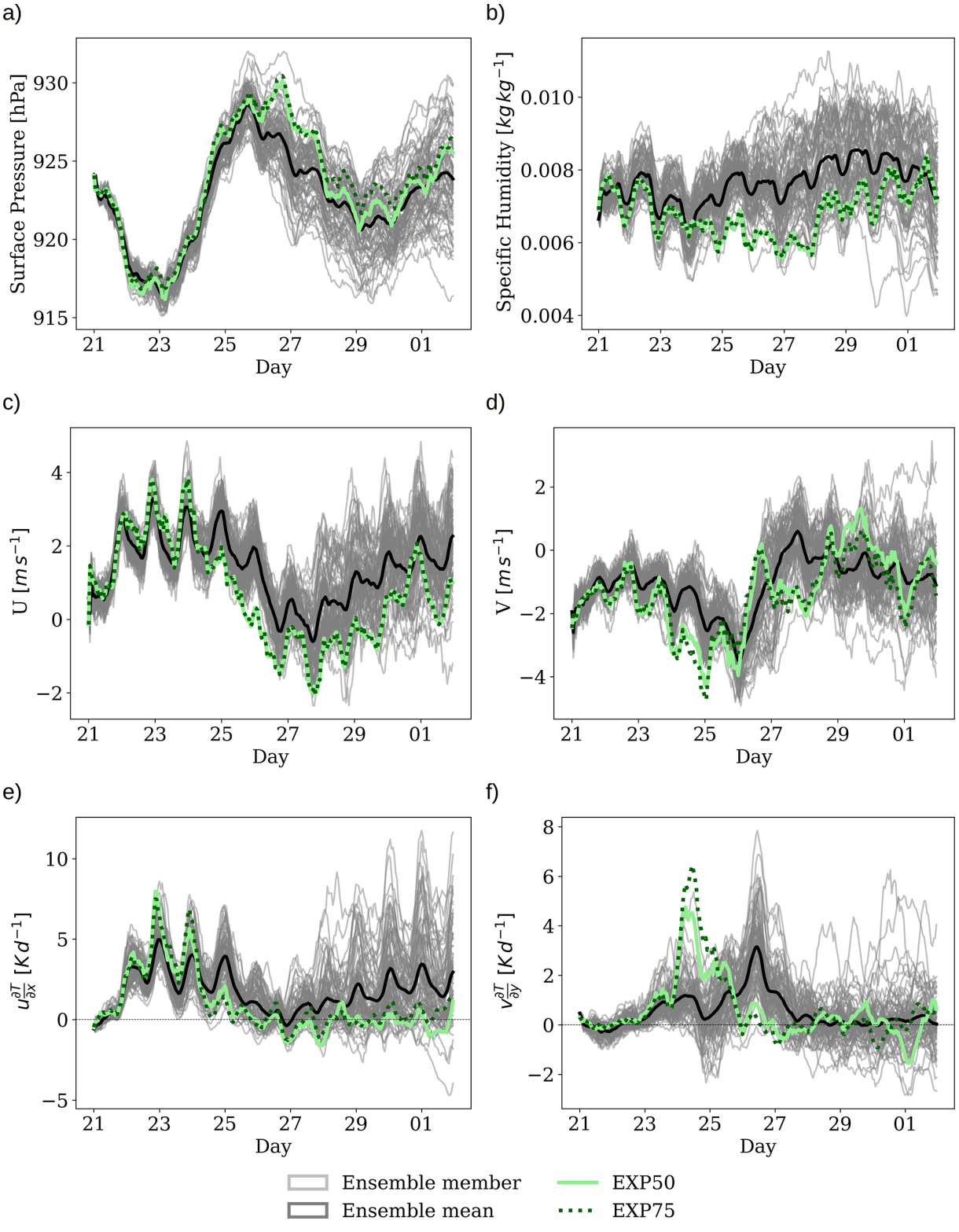


**Figure 6.** Time series of **(a)** surface pressure, **(b)** near-surface specific humidity, **(c)** $U$ component of wind, **(d)** $V$ component of wind, **(e, f)** temperature advection for each component at 1000 hPa. Data from the optimized trajectory are shown alongside the individual ensemble members (in gray) and the ensemble mean (thick black line).

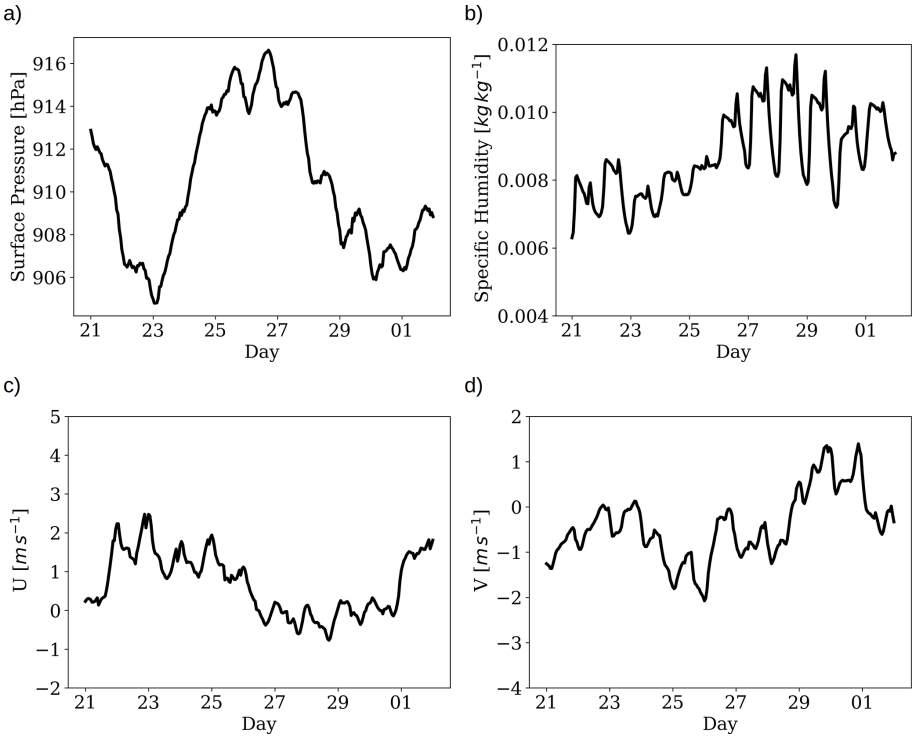


**Figure 7.** Same as Fig. 6 except using the ERA5 reanalysis data. Time series of **(a)** surface pressure **(b)** near-surface specific humidity **(c)** $U$ component of wind and **(d)** $V$ component of wind at 1000 hPa.

heatwave intensity by 3.7°C according to NeuralGCM. These results align with prior studies linking extreme heat to persistent blocking patterns (Screen and Simmonds, 2014; Mass et al., 2024). Specifically, the optimized geopotential height anomalies and spectrum reflect enhanced blocking dynamics with an amplification of wavenumbers 1–5. The resulting increase in near-surface temperature and 500 hPa geopotential heights has realistic features when compared to the ERA5 data, as can be seen in Fig. 5.

While NeuralGCM resolves large-scale dynamics, its omission of land-atmosphere feedbacks (e.g., soil moisture (Duan et al., 2025a)) likely results in a conservative estimate of heatwave amplification. For instance, soil moisture–temperature coupling is known to cause stronger heatwave persistence (e.g., (Suarez-Gutierrez et al., 2020; White et al., 2023)), implying that the model might underestimate extremes when such feedbacks are neglected. Additionally, the model's coarse horizontal resolution (2.8°) introduces biases in capturing localized extreme conditions associated with the PN2021 event. As shown by the higher resolution simulation, the use of finer grids (1.4°) allows for more accurate estimates of the extreme temperatures of the PN2021 event although the main dynamical changes introduced by the optimized solution remain similar to the coarse resolution simulation. In addition, while we have chosen to use NeuralGCM for this study, the method could be applied to any climate model that supports automatic differentiation. This includes all existing purely data-driven models, such as GraphCast, Pangu-Weather, FourCastNet, and FuXi. While data-driven models could provide faster simulations at higher horizontal resolutions, their dual-initial-condition requirement introduces ambiguity about finding optimal initial conditions. For instance, Vonich and Hakim (2024) optimized both inputs for GraphCast to reconstruct the 2021 heatwave, but this approach demands simultaneous perturbation of two distinct states. Validation against models like Pangu-Weather demonstrated that results remained consistent despite this added complexity, suggesting robustness in the dual-input framework. However, NeuralGCM's hybrid design simplifies the workflow by requiring only a single initial condition. In addition, Selz and Craig (2023) showed that data-driven forecast models represent poorly small-scale perturbations, often filtering them out, which could limit the applicability of the method. The extent to which this could affect hybrid models such as NeuralGCM, which make use of a traditional dynamical core, remains unclear.

The optimization process involves making several decisions and setting specific parameters. While we do not present the results here, we have explored a limited subset of the broader hyperparameter space – specifically, the learning rate ($\alpha$) and the loss function parameters ($\beta$, $\boldsymbol{\lambda}$). We found that, for the loss function parameters defined in Sect. 2, a large learning rate induces instability, causing substantial perturbations without a corresponding increase in

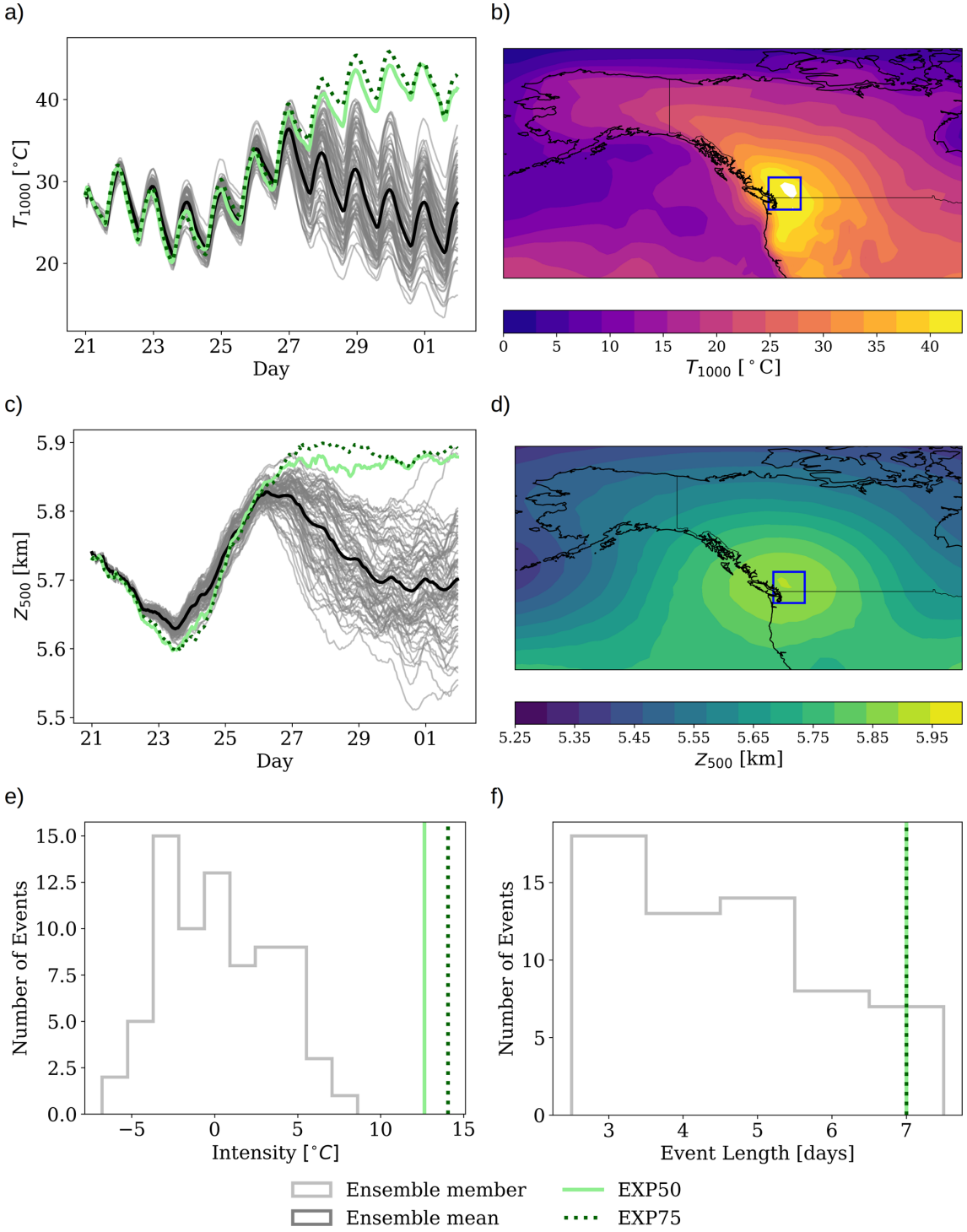


**Figure 8.** Same as Fig. 4 except with 1.4° resolution model. **(a)** Time series of 1000 hPa temperature for two optimized trajectories (50 and 75 steps) and a 75-member ensemble with its mean. **(b)** Spatial map of average temperature anomalies from the 75-step run. **(c)** Time series of 500 hPa geopotential height for the same set of simulations. **(d)** Spatial map of 500 hPa geopotential height anomalies during the event period. **(e, f)** Time series of heatwave intensity (defined by Eq. 4) and duration for the ensemble and optimized cases.

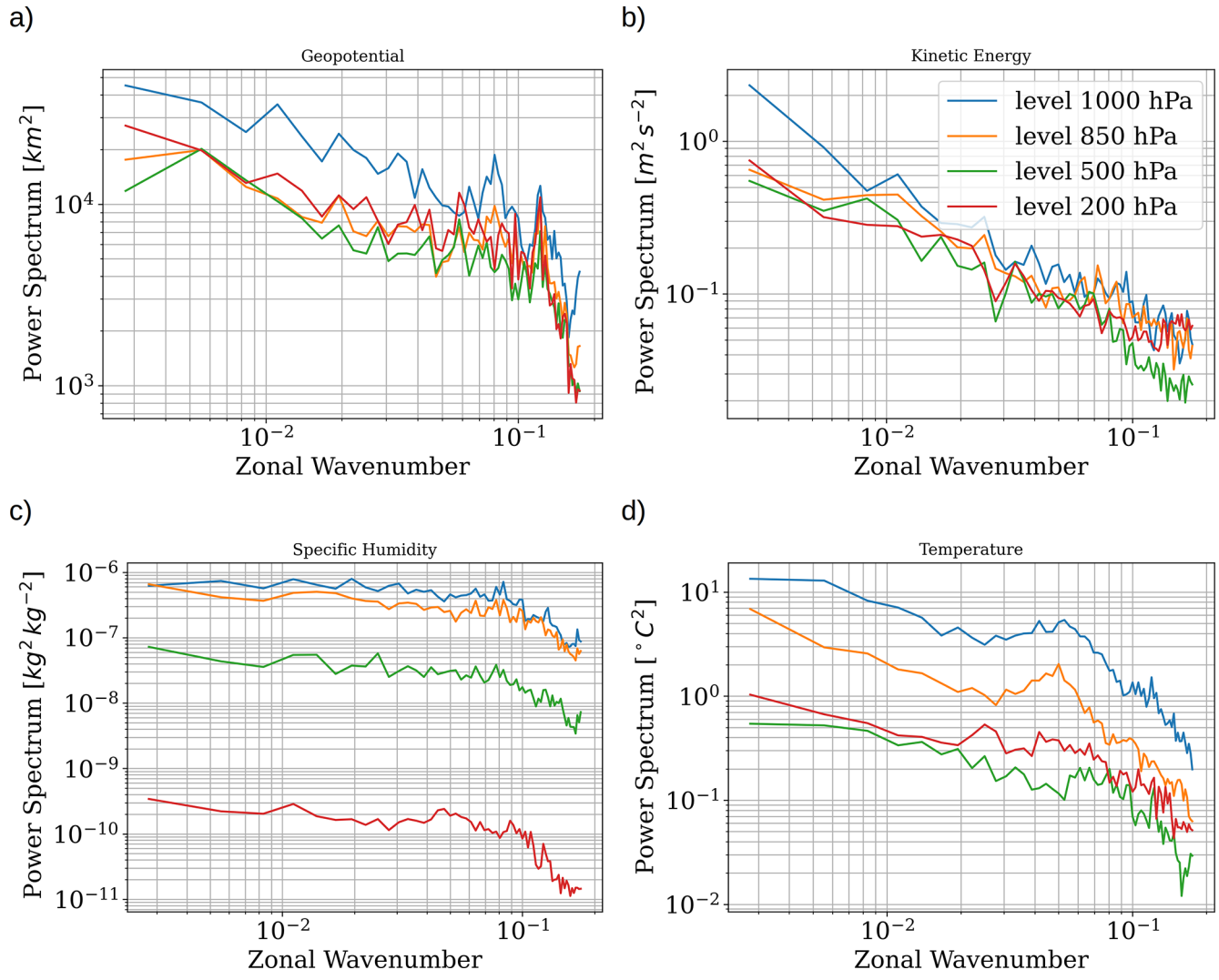


**Figure 9.** Power spectra of the three-dimensional perturbations at selected vertical levels for the geopotential **(a)**, kinetic energy **(b)**, specific humidity **(c)** and temperature **(d)**. Results are shown at the 1000, 850, 500 and 200 hPa vertical levels.

temperature and, in some cases, triggering numerical instabilities that cause the simulations to fail. To ensure stability, our tests showed that $\alpha$ must be on the order of $10^{-9}$ or smaller. This stability condition varies in a nonlinear and nontrivial manner with changes in $\boldsymbol{\lambda}$. Furthermore, the consistency of the results across the EXP50 and EXP75 experiments and the simulations at two different resolutions – all of which yield trajectories more extreme than the 75-member stochastic ensemble – suggests that the optimized perturbations are not simply initialization artifacts. However, a systematic quantification of the sensitivity to the initial state and a thorough exploration of the hyperparameter space lie beyond the scope of this study. Fine-tuning the parameters might allow for improved efficiency in computational cost, where a more extreme solution is found with a reduced number of optimization steps. In addition, we note that we have chosen to optimize the temperature field at the 1000 hPa pressure level, as the NeuralGCM converts from $\sigma$ coordinate levels to standard ERA5 pressure levels (Kochkov et al., 2024). Over regions with significant elevation, such as the Canadian Rockies, the 1000 hPa pressure level is often below the surface. Using the 1000 hPa temperature ($T_{1000}$) can therefore yield physically inconsistent values when optimizing for near-surface extreme events. To ensure the optimized initial conditions lead to physically meaningful and surface-relevant extreme temperatures across the entire domain, we included the time series and the spatial distribution of 850 hPa temperature ($T_{850}$) in the Supplementary Information. The results for $T_{1000}$ and $T_{850}$ appear to be consistent.

To evaluate the robustness and physical realism of the NeuralGCM-optimized initial conditions, these perturbations should be tested in a conventional numerical weather prediction model (e.g., Environment and Climate Change Canada's model; Buehner et al., 2015). Such cross-model validation

would reveal the universality of the results and help isolate NeuralGCM-specific biases. Additionally, running these scenarios in a fully physical model would explicitly account for land–atmosphere interactions and feedbacks, and assess whether the extreme trajectories persist under more detailed dynamics and physics.

Our method focuses on optimizing initial conditions, assuming the underlying model physics (whether learned or explicit) are fixed and skillful. An alternative approach could involve optimizing model parameters themselves (as done, for example, by Alet et al., 2025 to generate ensembles), though this would require careful regularization to ensure the resulting model remains physically plausible.

The computational efficiency of the ML and hybrid models coupled with their differentiable properties, opens avenues for exploring extreme events – from heatwaves to precipitation extremes and compound disasters. For example, a similar optimization problem could be formulated for StormCast (Pathak et al., 2024) to allow us to search for extreme precipitation events in an emulated regional climate model. One could also formulate the loss function such that the large deviation theory rate function is minimized, leading to "typical" trajectories of extremes (Grafke and Vanden-Eijnden, 2019; Zakine and Vanden-Eijnden, 2023). We could also envision loss functions with hard constraints on the perturbation which impose conservation laws as opposed to simply imposing small perturbations.

## 5 Conclusions

We introduce a differentiable-storyline framework that leverages automatic differentiation in hybrid climate models to directly optimize initial conditions and generate physically coherent extreme-heatwave trajectories at a fraction of the computational cost of traditional ensemble methods. For example, our 50-step optimization run produced a more extreme event than any member of a 75-member ensemble, while using 33 % less computational resources than it took to generate that ensemble. When applied to the PN2021 heatwave, our approach produces an intensification of nearly 3.7°C by isolating high-impact circulation patterns, specifically enhanced blocking and Rossby-wave, demonstrating both its dynamical fidelity and efficiency. While this proof-of-concept focuses on NeuralGCM and a single case study, the optimization paradigm is agnostic to model architecture and event type, offering a transformative tool for rapid, process-based risk assessment of diverse climate extremes in a warming world.

*Code availability.* The data generated and experiments can be reproduced using the following code: https://doi.org/10.5281/zenodo.15649393 (Whittaker, 2025).

*Data availability.* All data generated using the code and ERA5 data used in this study is openly available from the Copernicus Data Store at https://doi.org/10.24381/cds.143582cf (Hersbach et al., 2020).

*Supplement.* The supplement related to this article is available online at https://doi.org/10.5194/wcd-7-393-2026-supplement.

*Author contributions.* TW and ADL conceptualized the method. TW designed and performed the numerical experiments. TW and ADL prepared the manuscript with equal contribution.

*Competing interests.* The contact author has declared that neither of the authors has any competing interests.

*Disclaimer.* Publisher's note: Copernicus Publications remains neutral with regard to jurisdictional claims made in the text, published maps, institutional affiliations, or any other geographical representation in this paper. The authors bear the ultimate responsibility for providing appropriate place names. Views expressed in the text are those of the authors and do not necessarily reflect the views of the publisher.

*Acknowledgements.* Tim Whittaker would like to thank Robin Noyelle, and Eric Vanden-Eijnden for discussions which led to this project. Tim Whittaker would like to thank Seth Taylor and Tangui Picart for discussions and comments on the draft. The authors would like to thank Frédérik Toupin, Katja Winger, and Francois Roberge for maintaining a user-friendly local computing facility. This research was enabled in part by support provided by Calcul Québec (calculquebec.ca) and the Digital Research Alliance of Canada (alliancecan.ca). In addition, the authors would like to thank the two anonymous reviewers for the valuable input.

*Financial support.* This research has been supported by the Natural Sciences and Engineering Research Council of Canada (grant nos. RGPIN-2020-05631 and CCGSD-588387-2024).

*Review statement.* This paper was edited by Gwendal Rivière and reviewed by two anonymous referees.

## References

Alet, F., Price, I., El-Kadi, A., Masters, D., Markou, S., Andersson, T. R., Stott, J., Lam, R., Willson, M., Sanchez-Gonzalez, A., and Battaglia, P.: Skillful joint probabilistic weather forecasting from marginals, arXiv [preprint], https://doi.org/10.48550/arXiv.2506.10772, 12 June 2025.

Baño-Medina, J., Sengupta, A., Doyle, J. D., Reynolds, C. A., Watson-Parris, D., and Monache, L. D.: Are AI weather

models learning atmospheric physics? A sensitivity analysis of cyclone Xynthia, npj Clim. Atmos. Sci., 8, 92, https://doi.org/10.1038/s41612-025-00949-6, 2025.

Bartusek, S., Kornhuber, K., and Ting, M.: 2021 North American heatwave amplified by climate change-driven nonlinear interactions, Nat. Clim. Change, 12, 1143–1150, https://doi.org/10.1038/s41558-022-01520-4, 2022.

Bi, K., Xie, L., Zhang, H., Chen, X., Gu, X., and Tian, Q.: Accurate medium-range global weather forecasting with 3D neural networks, Nature, 619, 533–538, https://doi.org/10.1038/s41586-023-06185-3, 2023.

Bradbury, J., Frostig, R., Hawkins, P., Johnson, M. J., Leary, C., Maclaurin, D., Necula, G., Paszke, A., VanderPlas, J., Wanderman-Milne, S., and Zhang, Q.: JAX: composable transformations of Python+NumPy programs, GitHub [code], http://github.com/jax-ml/jax, 2018.

Buehner, M., McTaggart-Cowan, R., Beaulne, A., Charette, C., Garand, L., Heilliette, S., Lapalme, E., Laroche, S., Macpherson, S. R., Morneau, J., and Zadra, A.: Implementation of Deterministic Weather Forecasting Systems Based on Ensemble–Variational Data Assimilation at Environment Canada. Part I: The Global System, Mon. Weather Rev., 143, 2532–2559, https://doi.org/10.1175/MWR-D-14-00354.1, 2015.

Camps-Valls, G., Fernández-Torres, M.-Á., Cohrs, K.-H., Höhl, A., Castelletti, A., Pacal, A., Robin, C., Martinuzzi, F., Papoutsis, I., Prapas, I., Pérez-Aracil, J., Weigel, K., Gonzalez-Calabuig, M., Reichstein, M., Rabel, M., Giuliani, M., Mahecha, M. D., Popescu, O.-I., Pellicer-Valero, O. J., Ouala, S., Salcedo-Sanz, S., Sippel, S., Kondylatos, S., Happé, T., and Williams, T.: Artificial intelligence for modeling and understanding extreme weather and climate events, Nat. Commun., 16, 1919, https://doi.org/10.1038/s41467-025-56573-8, 2025.

Chen, L., Zhong, X., Zhang, F., Cheng, Y., Xu, Y., Qi, Y., and Li, H.: FuXi: a cascade machine learning forecasting system for 15-day global weather forecast, npj Clim. Atmos. Sci., 6, 190, https://doi.org/10.1038/s41612-023-00512-1, 2023.

Comeau, E., Di Luca, A., and Kirchmeier-Young, M.: From Mild to Extreme Heatwaves: Examining Trends in North America, Weather and Climate Extremes, 51, 100831, https://doi.org/10.1016/j.wace.2025.100831, 2025.

Conrick, R. and Mass, C. F.: The Influence of Soil Moisture on the Historic 2021 Pacific Northwest Heatwave, Mon. Weather Rev., 151, 1213–1228, https://doi.org/10.1175/MWR-D-22-0253.1, 2023.

DeepMind, Babuschkin, I., Baumli, K., Bell, A., Bhupatiraju, S., Bruce, J., Buchlovsky, P., Budden, D., Cai, T., Clark, A., Danihelka, I., Dedieu, A., Fantacci, C., Godwin, J., Jones, C., Hemsley, R., Hennigan, T., Hessel, M., Hou, S., Kapturowski, S., Keck, T., Kemaev, I., King, M., Kunesch, M., Martens, L., Merzic, H., Mikulik, V., Norman, T., Papamakarios, G., Quan, J., Ring, R., Ruiz, F., Sanchez, A., Sartran, L., Schneider, R., Sezener, E., Spencer, S., Srinivasan, S., Stanojević, M., Stokowiec, W., Wang, L., Zhou, G., and Viola, F.: The DeepMind JAX Ecosystem, GitHub, http://github.com/google-deepmind, 2020.

Deser, C., Lehner, F., Rodgers, K. B., Ault, T., Delworth, T. L., DiNezio, P. N., Fiore, A., Frankignoul, C., Fyfe, J. C., Horton, D. E., Kay, J. E., Knutti, R., Lovenduski, N. S., Marotzke, J., McKinnon, K. A., Minobe, S., Randerson, J., Screen, J. A., Simpson, I. R., and Ting, M.: Insights from Earth system model initial-condition large ensembles and future prospects, Nat. Clim. Change, 10, 277–286, https://doi.org/10.1038/s41558-020-0731-2, 2020.

Diffenbaugh, N. S. and Davenport, F. V.: On the impossibility of extreme event thresholds in the absence of global warming, Environ. Res. Lett., 16, 115014, https://doi.org/10.1088/1748-9326/ac2f1a, 2021.

Duan, S., Zhang, J., Bonfils, C., and Pallotta, G.: Testing NeuralGCM's capability to simulate future heatwaves based on the 2021 Pacific Northwest heatwave event, npj Clim. Atmos. Sci., 8, 251, https://doi.org/10.1038/s41612-025-01137-2, 2025a.

Duan, S. Q., McKinnon, K. A., and Simpson, I. R.: The Impact of Soil Preconditioning on the Evolution of Heatwaves Under Constrained Circulation: A Case Study of the 2021 Pacific Northwest Heatwave, Earth's Future, 13, e2025EF006216, https://doi.org/10.1029/2025EF006216, 2025b.

Ennis, K. E., Barnes, E. A., Arcodia, M. C., Fernandez, M. A., and Maloney, E. D.: Turning Up the Heat: Assessing 2-m Temperature Forecast Errors in AI Weather Prediction Models During Heat Waves, arXiv [preprint], https://doi.org/10.48550/arXiv.2504.21195, 29 April 2025.

Fischer, E. M., Beyerle, U., Bloin-Wibe, L., Gessner, C., Humphrey, V., Lehner, F., Pendergrass, A. G., Sippel, S., Zeder, J., and Knutti, R.: Storylines for unprecedented heatwaves based on ensemble boosting, Nat. Commun., 14, 4643, https://doi.org/10.1038/s41467-023-40112-4, 2023.

Gelbrecht, M., White, A., Bathiany, S., and Boers, N.: Differentiable programming for Earth system modeling, Geosci. Model Dev., 16, 3123–3135, https://doi.org/10.5194/gmd-16-3123-2023, 2023.

Gessner, C., Fischer, E. M., Beyerle, U., and Knutti, R.: Very Rare Heat Extremes: Quantifying and Understanding Using Ensemble Reinitialization, J. Climate, 34, 6619–6634, https://doi.org/10.1175/JCLI-D-20-0916.1, 2021.

Grafke, T. and Vanden-Eijnden, E.: Numerical computation of rare events via large deviation theory, Chaos, 29, 063118, https://doi.org/10.1063/1.5084025, 2019.

Hazeleger, W., van den Hurk, B. J. J. M., Min, E., van Oldenborgh, G. J., Petersen, A. C., Stainforth, D. A., Vasileiadou, E., and Smith, L. A.: Tales of future weather, Nat. Clim. Change, 5, 107–113, https://doi.org/10.1038/nclimate2450, 2015.

Hersbach, H., Bell, B., Berrisford, P., Hirahara, S., Horányi, A., Muñoz-Sabater, J., Nicolas, J., Peubey, C., Radu, R., Schepers, D., Simmons, A., Soci, C., Abdalla, S., Abellan, X., Balsamo, G., Bechtold, P., Biavati, G., Bidlot, J., Bonavita, M., De Chiara, G., Dahlgren, P., Dee, D., Diamantakis, M., Dragani, R., Flemming, J., Forbes, R., Fuentes, M., Geer, A., Haimberger, L., Healy, S., Hogan, R. J., Hólm, E., Janisková, M., Keeley, S., Laloyaux, P., Lopez, P., Lupu, C., Radnoti, G., de Rosnay, P., Rozum, I., Vamborg, F., Villaume, S., and Thépaut, J.-N.: The ERA5 global reanalysis, Q. J. Roy. Meteor. Soc., 146, 1999–2049, https://doi.org/10.1002/qj.3803, 2020.

Kochkov, D., Yuval, J., Langmore, I., Norgaard, P., Smith, J., Mooers, G., Klöwer, M., Lottes, J., Rasp, S., Düben, P., Hatfield, S., Battaglia, P., Sanchez-Gonzalez, A., Willson, M., Brenner, M. P., and Hoyer, S.: Neural general circulation models for weather and climate, Nature, 632, 1060-1066, https://doi.org/10.1038/s41586-024-07744-y, 2024.

Kurth, T., Subramanian, S., Harrington, P., Pathak, J., Mardani, M., Hall, D., Miele, A., Kashinath, K., and Anandkumar, A.: FourCastNet: Accelerating Global High-Resolution Weather Forecasting Using Adaptive Fourier Neural Operators, 1–11 pp., https://doi.org/10.1145/3592979.3593412, 2023.

Lam, R., Sanchez-Gonzalez, A., Willson, M., Wirnsberger, P., Fortunato, M., Alet, F., Ravuri, S., Ewalds, T., Eaton-Rosen, Z., Hu, W., Merose, A., Hoyer, S., Holland, G., Vinyals, O., Stott, J., Pritzel, A., Mohamed, S., and Battaglia, P.: Learning skillful medium-range global weather forecasting, Science, 382, 1416–1421, https://doi.org/10.1126/science.adi2336, 2023.

Loikith, P. C. and Kalashnikov, D. A.: Meteorological Analysis of the Pacific Northwest June 2021 Heatwave, Mon. Weather Rev., 151, 1303–1319, https://doi.org/10.1175/MWR-D-22-0284.1, 2023.

Lorenz, E.: Predictability – a problem partly solved, Predictability of Weather and Climate, Cambridge University Press, edited by: Palmer, T. and Hagedorn, R., 4–8 September 1995, pp. 40–58, 2006.

Lucarini, V., Melinda Galfi, V., Riboldi, J., and Messori, G.: Typicality of the 2021 Western North America summer heatwave, Environ. Res. Lett., 18, 015004, https://doi.org/10.1088/1748-9326/acab77, 2023.

Maher, N., Milinski, S., and Ludwig, R.: Large ensemble climate model simulations: introduction, overview, and future prospects for utilising multiple types of large ensemble, Earth Syst. Dynam., 12, 401–418, https://doi.org/10.5194/esd-12-401-2021, 2021.

Mahesh, A., Collins, W., Bonev, B., Brenowitz, N., Cohen, Y., Elms, J., Harrington, P., Kashinath, K., Kurth, T., North, J., OBrien, T., Pritchard, M., Pruitt, D., Risser, M., Subramanian, S., and Willard, J.: Huge Ensembles Part I: Design of Ensemble Weather Forecasts using Spherical Fourier Neural Operators, arXiv [preprint], https://doi.org/10.48550/arXiv.2408.03100, 6 August 2024a.

Mahesh, A., Collins, W., Bonev, B., Brenowitz, N., Cohen, Y., Harrington, P., Kashinath, K., Kurth, T., North, J., OBrien, T., Pritchard, M., Pruitt, D., Risser, M., Subramanian, S., and Willard, J.: Huge Ensembles Part II: Properties of a Huge Ensemble of Hindcasts Generated with Spherical Fourier Neural Operators, arXiv [preprint], https://doi.org/10.48550/arXiv.2408.01581, 2 August 2024b.

Manshausen, P., Cohen, Y., Pathak, J., Pritchard, M., Garg, P., Mardani, M., Kashinath, K., Byrne, S., and Brenowitz, N.: Generative Data Assimilation of Sparse Weather Station Observations at Kilometer Scales, arXiv [preprint], https://doi.org/10.48550/arXiv.2406.16947, 2024.

Mass, C., Ovens, D., Christy, J., and Conrick, R.: The Pacific Northwest Heat Wave of 25–30 June 2021: Synoptic/Mesoscale Conditions and Climate Perspective, Weather Forecast., 39, 275–291, https://doi.org/10.1175/WAF-D-23-0154.1, 2024.

Materia, S., García, L. P., van Straaten, C., O, S., Mamalakis, A., Cavicchia, L., Coumou, D., de Luca, P., Kretschmer, M., and Donat, M.: Artificial intelligence for climate prediction of extremes: State of the art, challenges, and future perspectives, WIREs Clim. Change, 15, e914, https://doi.org/10.1002/wcc.914, 2024.

Mo, R., Lin, H., and Vitart, F.: An anomalous warm-season trans-Pacific atmospheric river linked to the 2021 western North America heatwave, Communications Earth & Environment, 3, 127, https://doi.org/10.1038/s43247-022-00459-w, 2022.

Neal, E., Huang, C. S. Y., and Nakamura, N.: The 2021 Pacific Northwest Heat Wave and Associated Blocking: Meteorology and the Role of an Upstream Cyclone as a Diabatic Source of Wave Activity, Geophys. Res. Lett., 49, e2021GL097699, https://doi.org/10.1029/2021GL097699, 2022.

Noyelle, R., Caubel, A., Meurdesoif, Y., Yiou, P., and Faranda, D.: Statistical and dynamical aspects of extremely hot summers in Western Europe sampled with a rare events algorithm, J. Climate, https://doi.org/10.1175/JCLI-D-24-0635.1, 2025.

Pasche, O. C., Wider, J., Zhang, Z., Zscheischler, J., and Engelke, S.: Validating Deep Learning Weather Forecast Models on Recent High-Impact Extreme Events, Artificial Intelligence for the Earth Systems, 4, e240033, https://doi.org/10.1175/AIES-D-24-0033.1, 2025.

Pathak, J., Subramanian, S., Harrington, P., Raja, S., Chattopadhyay, A., Mardani, M., Kurth, T., Hall, D., Li, Z., Azizzadenesheli, K., Hassanzadeh, P., Kashinath, K., and Anandkumar, A.: FourCastNet: A Global Data-driven High-resolution Weather Model using Adaptive Fourier Neural Operators, arXiv [preprint], https://doi.org/10.48550/arXiv.2202.11214, 22 February 2022.

Pathak, J., Cohen, Y., Garg, P., Harrington, P., Brenowitz, N., Durran, D., Mardani, M., Vahdat, A., Xu, S., Kashinath, K., and Pritchard, M.: Kilometer-Scale Convection Allowing Model Emulation using Generative Diffusion Modeling, arXiv [preprint], https://doi.org/10.48550/arXiv.2408.10958, 20 August 2024.

Philip, S. Y., Kew, S. F., van Oldenborgh, G. J., Anslow, F. S., Seneviratne, S. I., Vautard, R., Coumou, D., Ebi, K. L., Arrighi, J., Singh, R., van Aalst, M., Pereira Marghidan, C., Wehner, M., Yang, W., Li, S., Schumacher, D. L., Hauser, M., Bonnet, R., Luu, L. N., Lehner, F., Gillett, N., Tradowsky, J. S., Vecchi, G. A., Rodell, C., Stull, R. B., Howard, R., and Otto, F. E. L.: Rapid attribution analysis of the extraordinary heat wave on the Pacific coast of the US and Canada in June 2021, Earth Syst. Dynam., 13, 1689–1713, https://doi.org/10.5194/esd-13-1689-2022, 2022.

Plotkin, D. A., Webber, R. J., O'Neill, M. E., Weare, J., and Abbot, D. S.: Maximizing Simulated Tropical Cyclone Intensity With Action Minimization, J. Adv. Model. Earth Sy., 11, 863–891, https://doi.org/10.1029/2018MS001419, 2019.

Ragone, F. and Bouchet, F.: Rare Event Algorithm Study of Extreme Warm Summers and Heatwaves Over Europe, Geophys. Res. Lett., 48, e2020GL091197, https://doi.org/10.1029/2020GL091197, 2021.

Ragone, F., Wouters, J., and Bouchet, F.: Computation of extreme heat waves in climate models using a large deviation algorithm, P. Natl. Acad. Sci. USA, 115, 24–29, https://doi.org/10.1073/pnas.1712645115, 2018.

Rasp, S., Hoyer, S., Merose, A., Langmore, I., Battaglia, P., Russell, T., Sanchez-Gonzalez, A., Yang, V., Carver, R., Agrawal, S., Chantry, M., Ben Bouallegue, Z., Dueben, P., Bromberg, C., Sisk, J., Barrington, L., Bell, A., and Sha, F.: WeatherBench 2: A Benchmark for the Next Generation of Data-Driven Global Weather Models, Journal of Advances in Modeling Earth Sy., 16, e2023MS004019, https://doi.org/10.1029/2023MS004019, 2024.

Schumacher, D. L., Hauser, M., and Seneviratne, S. I.: Drivers and Mechanisms of the 2021 Pacific Northwest Heatwave, Earth's Future, 10, e2022EF002967, https://doi.org/10.1029/2022EF002967, 2022.

Screen, J. A. and Simmonds, I.: Amplified mid-latitude planetary waves favour particular regional weather extremes, Nat. Clim. Change, 4, 704–709, https://doi.org/10.1038/nclimate2271, 2014.

Selz, T. and Craig, G. C.: Can Artificial Intelligence-Based Weather Prediction Models Simulate the Butterfly Effect?, Geophys. Res. Lett., 50, e2023GL105747, https://doi.org/10.1029/2023GL105747, 2023.

Shepherd, T. G.: Storyline approach to the construction of regional climate change information, P. Roy. Soc. A-Math. Phy., 475, 20190013, https://doi.org/10.1098/rspa.2019.0013, 2019.

Sillmann, J., Shepherd, T. G., van den Hurk, B., Hazeleger, W., Martius, O., Slingo, J., and Zscheischler, J.: Event-Based Storylines to Address Climate Risk, Earth's Future, 9, e2020EF001783, https://doi.org/10.1029/2020EF001783, 2021.

Solvik, K., Penny, S. G., and Hoyer, S.: 4D-Var Using Hessian Approximation and Backpropagation Applied to Automatically Differentiable Numerical and Machine Learning Models, J. Adv. Model. Earth Sy., 17, e2024MS004608, https://doi.org/10.1029/2024MS004608, 2025.

Suarez-Gutierrez, L., Müller, W. A., Li, C., and Marotzke, J.: Dynamical and thermodynamical drivers of variability in European summer heat extremes, Clim. Dynam., 54, 4351–4366, https://doi.org/10.1007/s00382-020-05233-2, 2020.

Vonich, P. T. and Hakim, G. J.: Predictability Limit of the 2021 Pacific Northwest Heatwave From Deep-Learning Sensitivity Analysis, Geophys. Res. Lett., 51, e2024GL110651, https://doi.org/10.1029/2024GL110651, 2024.

Webber, R. J., Plotkin, D. A., O'Neill, M. E., Abbot, D. S., and Weare, J.: Practical rare event sampling for extreme mesoscale weather, Chaos, 29, 053109, https://doi.org/10.1063/1.5081461, 2019.

White, R. H., Anderson, S., Booth, J. F., Braich, G., Draeger, C., Fei, C., Harley, C. D. G., Henderson, S. B., Jakob, M., Lau, C.-A., Mareshet Admasu, L., Narinesingh, V., Rodell, C., Roocroft, E., Weinberger, K. R., and West, G.: The unprecedented Pacific Northwest heatwave of June 2021, Nat. Commun., 14, 727, https://doi.org/10.1038/s41467-023-36289-3, 2023.

Whittaker, T.: timwhittaker/ExtremeStorylines, Zenodo [code], https://doi.org/10.5281/zenodo.15649394, 2025.

Yiou, P. and Jézéquel, A.: Simulation of extreme heat waves with empirical importance sampling, Geosci. Model Dev., 13, 763–781, https://doi.org/10.5194/gmd-13-763-2020, 2020.

Zakine, R. and Vanden-Eijnden, E.: Minimum-Action Method for Nonequilibrium Phase Transitions, Phys. Rev. X, 13, 041044, https://doi.org/10.1103/PhysRevX.13.041044, 2023.

Zhang, Z., Fischer, E., Zscheischler, J., and Engelke, S.: Numerical models outperform AI weather forecasts of record-breaking extremes, arXiv [preprint], https://doi.org/10.48550/arXiv.2508.15724, 21 August 2025.