Supporting Information for "Revealing the statistics of extreme events hidden in short weather forecast data "

Justin Finkel[1], Edwin P. Gerber[2], Dorian S. Abbot[3], Jonathan Weare[2]

[1]Committee on Computational and Applied Mathematics, University of Chicago

[2]Courant Institute of Mathematical Sciences, New York University

[3]Department of Geophysical Sciences, University of Chicago

Contents of this file

Introduction

Our work relies completely on publicly available datasets of reanalysis and hindcasts, which we describe in the subsequent section. We then lay out the numerical procedure to compute rates and seasonal distributions using transition path theory (TPT). We then

present the formulas used to display results in the main text. Finally, we document the method used to select parameters. A software implementation will be made available in a public repository at the time of publication.

Dataset description

We use two different datasets for this study.

• ERA-5: reanalysis product from ECMWF (Hersbach et al., 2020), spanning 1959-2019. We used daily averaged temperature, geopotential height, zonal and meridional wind at 10, 100, 500, and 850 hPa levels at a resolution of 2.5°. ERA-5 was downloaded from the Copernicus Data Store `https://cds.climate.copernicus.eu/`

• S2S: perturbed reforecast (hindcast) ensembles from the 2017 model version of the ECMWF IFS, launched every Monday and Thursday, from fall 1996 through spring 2016, at the same resolution as ERA5 above. Each hindcast ensemble has 10 perturbed members which run for 47 day, including the initialization date. We used the same time and space resolution as with ERA5. S2S was downloaded from the ECMWF data portal `https://ecmwf.int`.

The S2S dataset can be summarized as follows:

$$\text{S2S dataset} = \{\mathbf{X}_i(t) : i = 1, \ldots, N\} \tag{1}$$

Here, $\mathbf{X}$ denotes the state vector of all relevant meteorological variables. Associated with each calendar day t is a subset of "active" indices, $\mathcal{I}(t)$, containing the trajectories launched sometime between $t - 46$ and t. The same notation can be used for the zonal-mean zonal wind itself, i.e., $U_i(t)$ as a function of $\mathbf{X}_i(t)$.

1. MSM calculations

The main text explains the computation of the committor, q_t^+, and probability distribution, π_t, for a Markov chain. Here we further explain how to compute statistics of the hitting time,

$$\tau_t^+ = \min\{s \geq 0 : (t+s, \mathbf{X}(t+s)) \in B\} \tag{2}$$

1.0.1. PMF of hitting time

The probability mass function of τ_t^+ conditioned on an initial set $S_{t,j}$ can be computed with a recursion relation similar to that for q_t^+. As a base case, we observe that only way for $\tau_t^+ = 0$ is for the system to already be in state B:

$$\mathbb{P}\{\tau_t^+ = 0 | \mathbf{X}(t) \in j\} = \begin{cases} 1 & \text{if } (t,j) \in B \\ 0 & \text{otherwise} \end{cases} \tag{3}$$

For the case $\tau_t^+ \geq 1$, the hitting time is one plus the hitting time at the next step:

$$\mathbb{P}\{\tau_t^+ = s | \mathbf{X}(t) \in j\} = \sum_k P_{t,t+1}(j,k) \mathbb{P}\{\tau_{t+1}^+ = s-1 | \mathbf{X}(t+1) \in S_{t+1,k}\} \tag{4}$$

To carry out this recursion, we compute the left-hand side for all t and j with a fixed s before incrementing $s \to s+1$. After doing this for $0, 1, \ldots, s$, we then have a *time-limited* committor,

$$q_{t,\sigma}^+(j) = \mathbb{P}\{\tau_t^+ \leq \sigma | \mathbf{X}(t) \in S_{t,j}\} = \sum_{s=0}^{\sigma} \mathbb{P}\{\tau_t^+ = s | \mathbf{X}(t) \in S_{t,j}\}, \tag{5}$$

which is the preferred version of the committor for some other studies, e.g., (Lucente et al., 2022). Below we use the time-limited committor for hyperparameter tuning.

1.0.2. Expectation of hitting time

The most intricate computation is that of the expected lead time $\eta_t^+ = \mathbb{E}[\tau_t^+ | \tau_t^+ + t < T_1]$. We do this via the *moment-generating function*,

$$M_t(j;\lambda) = \mathbb{E}\big[\exp\big(\lambda\tau_t^+\big)\mathbb{I}\{\tau_t^+ + t < T_1\}\big|\mathbf{X}(t) \in S_j\big] \tag{6}$$

Along a single trajectory, we have $\tau_t^+ = \tau_{t+1}^+ + 1$ and thus $\exp\big(\lambda\tau_t^+\big) = \exp\big(\lambda\tau_{t+1}^+\big)\exp\big(\lambda\big)$. Using this fact, there is a recursive relationship between M_t and M_{t+1}:

$$M_t(j;\lambda) = \begin{cases} \sum_k P_{t,t+1}(j,k)e^{\lambda}M_{t+1}(k;\lambda) & (t,j) \notin A \cup B \\ 1 & (t,j) \in B \\ 0 & (t,j) \in A \end{cases} \tag{7}$$

One can observe that $M(t;\lambda = 0)$ is the committor itself, q_t^+. But to get at τ_t^+, we now must differentiate with respect to λ:

$$\frac{\partial}{\partial\lambda}M_t(j;\lambda) = \begin{cases} \sum_k P_{t,t+1}(j,k)e^{\lambda}\Big[M_{t+1}(k,\lambda) + \frac{\partial}{\partial\lambda}M_{t+1}(k,\lambda)\Big] & (t,j) \notin A \cup B \\ 0 & (t,j) \in B \\ 0 & (t,j) \in A \end{cases} \tag{8}$$

The lead time can be expressed

$$\eta_t^+(j) = \frac{\mathbb{E}[e^{\lambda\tau_t^+}\mathbb{I}\{\tau_t^+ + t < T_t\}|\mathbf{X}(t) \in S_j]}{\mathbb{P}\{\tau_t^+ + t < T_1|\mathbf{X}(t) \in S_j\}} = \frac{[\partial M_t(j;\lambda)/\partial\lambda]_{\lambda=0}}{q_t^+(j)} \tag{9}$$

Therefore, the recursion relation for η_t^+ is found by setting $\lambda = 0$ in the recursion relation for $M_t(j;\lambda)$:

$$\eta_t^+(t,j) = \begin{cases} \frac{1}{q_t^+(j)}\sum_k P_{t,t+1}(j,k)e^{\lambda}q_{t+1}(k)\big[1 + \eta_{t+1}(k)\big] & (t,j) \notin A \cup B \\ 0 & (t,j) \in B \\ \text{undefined} & (t,j) \in A \end{cases} \tag{10}$$

The formulas for q^+ and η^+ are exact, straightforward to implement, and fast to compute after a Markov chain is constructed. How to choose parameters to construct the chain in the first place is the focus of the next section.

2. MSM hyperparameter selection

The MSM procedure uses two key parameters: the number of time delays δ, and the number of clusters to use in the resulting $(\delta+1)$-dimensional feature space. We optimize this choice by comparing the time-limited committor with $\sigma = 20$ according to the MSM ($q^{(\mathrm{MSM})}$, which takes on continuous values between 0 and 1) with that according to the S2S data directly ($q^{(\mathrm{S2S})}$, which takes on discrete values of 0 or 1). Forgive the temporary abuse of sub- and superscripts on q in this subsection. For any ensemble member $\mathbf{X}_i(t)$ at some time t between 0 and 46, the latter can be assessed directly by asking whether $\mathbf{X}_i$ achieves SSW before a time σ has elapsed. Fig. S1 shows $q^+_{t,\sigma}$ as a function of t and $U^{(\mathrm{th})}$ for $\sigma = 20$ days and two different zonal wind thresholds, 0 m/s (top) and -16 m/s (bottom). Black contours are level sets of the climatological probability density, π. The two columns show $q^+_{t,\sigma}$ estimated from S2S directly (left) and the MSM (right), and they match approximately by eye. To quantify their agreement, we use the log-likelihood commonly used for logistic regression, which is equivalent to the (negative) cross entropy of Bernoulli$\big(q^{(\mathrm{MSM})}\big)$ relative to Bernoulli$\big(q^{(\mathrm{S2S})}\big)$:

$$\mathrm{LL}\big(q^{(\mathrm{S2S})}, q^{(\mathrm{MSM})}\big) = \mathrm{mean}\Big\{q_i^{(\mathrm{S2S})} \log\Big(q_i^{(\mathrm{MSM})}\Big) + \Big(1 - q_i^{(\mathrm{S2S})}\Big) \log\Big(1 - q_i^{(\mathrm{MSM})}\Big)\Big\}. \quad (11)$$

The committors are evaluated at day 23 of the member i, and the mean is taken over the ensemble members i such that $q_i^{(\mathrm{MSM})}$ is strictly between zero and one on the 23rd day. We choose 23 days because (i) it is beyond the 15-day maximum lag time, before which not all time-lagged features are defined for $\delta = 15$, and (ii) it leaves room for 20 days of extra lead time for validation purposes. Other choices in this neighborhood do not affect our results appreciably.

For extremely rare events—say, with 1% probability, so that the sample mean $\overline{q}^{(\mathrm{S2S})} = 0.01$—there is a class imbalance problem: $q^{(\mathrm{MSM})}$ can achieve a high LL score with a "climatological" forecast of $q_i^{(\mathrm{MSM})} \equiv \overline{q}^{(\mathrm{S2S})}$. LL can still be used as a relative score to discriminating between parameter choices at a fixed $\overline{q}^{(\mathrm{S2S})}$, but we wish to choose MSM parameters that are optimal on average across a wide range of $U^{(\mathrm{th})}$ and therefore of $\overline{q}^{(\mathrm{S2S})}$. To compare these scales meaningfully, we convert LL into an absolute scale following (Benedetti, 2010; Miloshevich et al., 2022):

$$\text{Normalized skill} = \frac{\mathrm{LL}\big(q^{(\mathrm{S2S})}, q^{(\mathrm{MSM})}\big) - \mathrm{LL}\big(q^{(\mathrm{S2S})}, \overline{q}^{(\mathrm{S2S})}\big)}{\mathrm{LL}\big(q^{(\mathrm{S2S})}, q^{(\mathrm{S2S})}\big) - \mathrm{LL}\big(q^{(\mathrm{S2S})}, \overline{q}^{(\mathrm{S2S})}\big)} \tag{12}$$

The numerator is the improvement relative to the climatological forecast, and the denominator is the maximum possible improvement, when each $q_i^{(\mathrm{S2S})}$ is predicted exactly. Hence the normalized skill is always less than one, and typically above zero. The right-hand panels of Fig. S1 display the average normalized skill across thresholds (top) and the normalized skill as a function of threshold (bottom) for a few parameter choices. Increasing the number of clusters seems to help regardless of the number of delays. On the other hand, more delays don't always help; beyond $\delta = 5$ days, the dimensionality may be introducing degeneracies. Among the parameter choices considered, the top and bottom panel both suggest that $\delta = 5$ delays and $M_t = 150$ clusters achieves the best performance among all the MSM choices shown.

Fig. S2 displays further confirmation that this choice is reasonable and robust. We have repeated the entire MSM pipeline with 20 different random subsets of the years 1996-2015 (sampled without replacement), and plotted the corresponding rates as light purple curves in Fig. S2a. The resampled estimates cluster about the dashed curve, which

uses all 20 years simultaneously, for all but the most extreme events. On the other hand, in Fig. S2b, we used 15 delays and 200 clusters, which gave a slightly better normalized skill score (not shown). However, the resampled rates are systematically biased away from the full 20-year estimate, which raises the suspicion of overfitting. For this reason, we stick to the more modest choice of 5 delays and 150 clusters. However, all the qualitative results shown—the committors, expected lead times, and rate curves—remain remarkably consistent with different choices of clusters and delays. This gives us confidence in the MSM results shown in the main text.

References

Benedetti, R. (2010). Scoring rules for forecast verification. *Monthly Weather Review*, *138*(1), 203 - 211. Retrieved from `https://journals.ametsoc.org/view/journals/mwre/138/1/2009mwr2945.1.xml` doi: 10.1175/2009MWR2945.1

Hersbach, H., Bell, B., Berrisford, P., Hirahara, S., Horányi, A., Muñoz-Sabater, J., ... Thépaut, J.-N. (2020). The era5 global reanalysis. *Quarterly Journal of the Royal Meteorological Society*, *146*(730), 1999-2049. Retrieved from `https://rmets.onlinelibrary.wiley.com/doi/abs/10.1002/qj.3803` doi: https://doi.org/10.1002/qj.3803

Lucente, D., Rolland, J., Herbert, C., & Bouchet, F. (2022). Coupling rare event algorithms with data-based learned committor functions using the analogue markov chain. *Journal of Statistical Mechanics: Theory and Experiment*, *2022*(8), 083201.

Miloshevich, G., Cozian, B., Abry, P., Borgnat, P., & Bouchet, F. (2022). Probabilistic forecasts of extreme heatwaves using convolutional neural networks in a regime of

lack of data. *arXiv preprint arXiv:2208.00971*.

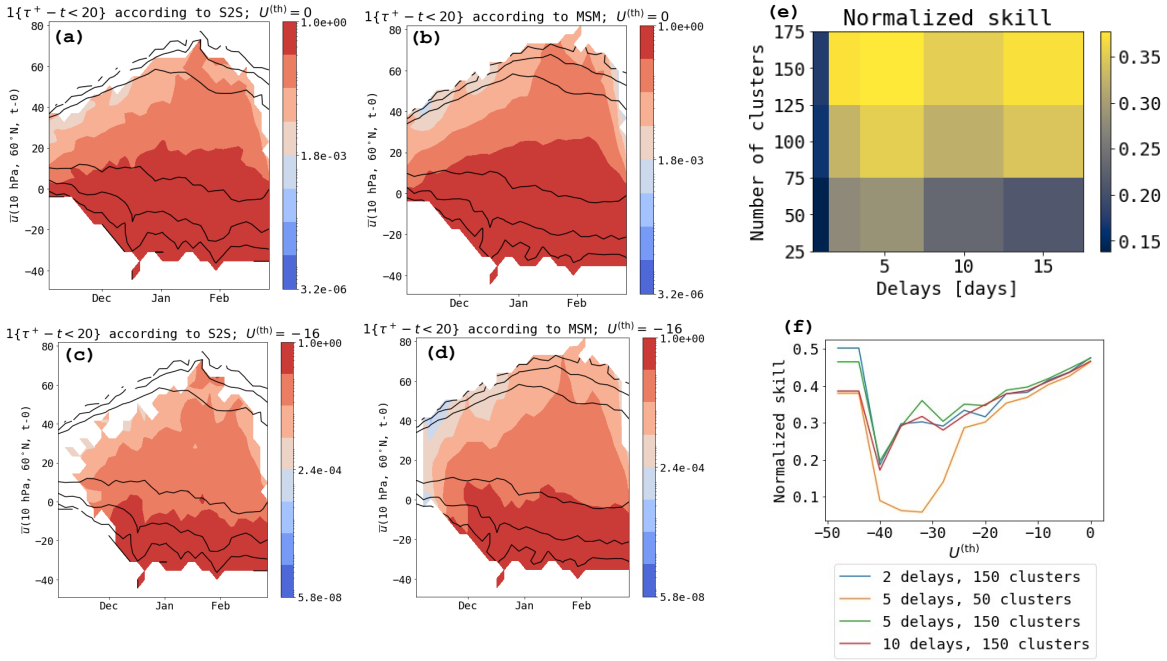

Figure S1. Time-limited committors for hyperparameter selection. Left two columns: shading indicates time-limited committor probabilities according to counting S2S hindcasts (left) and the MSM (right), at two different thresholds: $U^{(th)} = 0$ m/s (top) and -16 m/s (bottom). Black contours delineate level sets of the climatological probability density, derived from the same two methods. The MSM was constructed using 5 time delays and 150 clusters, a choice informed by systematic evaluation of the MSMs performance. The upper right panel shows the normalized skill of the MSM averaged over zonal wind thresholds, while the lower right panel plots the skill against thresholds for a few selected parameter sets.

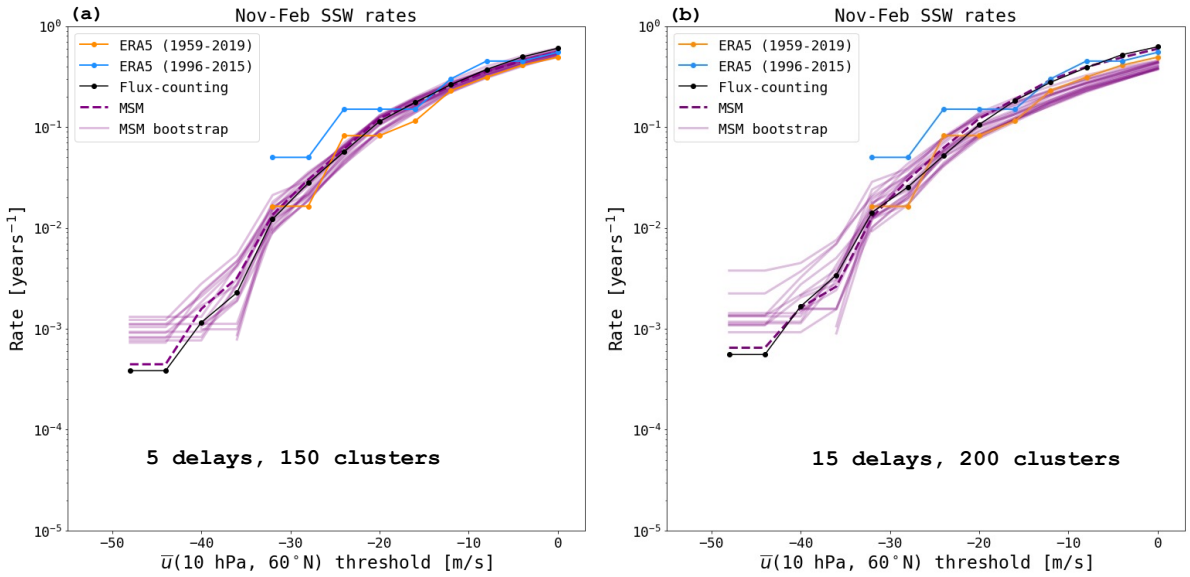

Figure S2. Bootstrap-resampled rates. Left: reproduction of Fig. 2, plus 20 additional rate curves obtained by resampling the 20 years without into random subsets of size 10, without replacement. Right: same for 15 delays and 200 clusters. The displacement of the bootstrapped curves from the main curve is a symptom of overfitting.