

Seasonal forecasting of European heatwaves using a feature selection framework

Corresponding Author: Dr Ronan McAdam

This file contains all editorial decision letters in order by version, followed by all author rebuttals in order by version.

Version 0:

Decision Letter:

**** Please ensure you delete the link to your author home page in this e-mail if you wish to forward it to your coauthors ****

Dear Dr McAdam,

Your manuscript titled "Feature selection for data-driven seasonal forecasts of European heatwaves" has now been seen by 3 reviewers, whose comments are appended below. You will see that they find your work of some potential interest. However, they have raised quite substantial concerns that must be addressed. In light of these comments, we cannot accept the manuscript for publication, but would be interested in considering a revised version that fully addresses these serious concerns. Specifically, we ask you to:

1. Fully justify methodological choices used for model construction and validation and provide more detailed explanations
2. Clarify the logic and advantages of proposed feature selection method over other methods and discuss associated physical mechanisms

We hope you will find the reviewers' comments useful as you decide how to proceed. Should additional work allow you to address these criticisms, we would be happy to look at a substantially revised manuscript. If you choose to take up this option, please either highlight all changes in the manuscript text file, or provide a list of the changes to the manuscript with your responses to the reviewers.

When resubmitting, please provide a point-by-point response to the reviewers' comments. Please submit your responses as a separate file, distinct from your cover letter where you can add responses to the Editors' comments that you do not want to be made available to the reviewers. Word files are preferred. We recommend that any figures, tables or graphs that are included in the response to reviewers are also included in the main article or Supplementary Information.

Please bear in mind that we will be reluctant to approach the reviewers again in the absence of substantial revisions.

If the revision process takes significantly longer than three months, we will be happy to reconsider your paper at a later date, as long as nothing similar has been accepted for publication at Communications Earth & Environment or published elsewhere in the meantime.

We are committed to providing a fair and constructive peer-review process. Please do not hesitate to contact us if you wish to discuss the revision in more detail.

Please use the following link to submit your revised manuscript, point-by-point response to the reviewers' comments with a list of your changes to the manuscript text (which should be in a separate document to any cover letter), a tracked-changes version of the manuscript (as a PDF file) and any completed checklist:

Link Redacted

**** This url links to your confidential home page and associated information about manuscripts you may have submitted or be reviewing for us. If you wish to forward this email to co-authors, please delete the link to your homepage first ****

Please do not hesitate to contact us if you have any questions or would like to discuss the required revisions further. Thank you for the opportunity to review your work.

Best regards,

Prof Seung-Ki Min
Editorial Board Member
Communications Earth & Environment
0000-0002-6749-010X

Alienor Lavergne, PhD
Senior Editor
Communications Earth & Environment

EDITORIAL POLICIES AND FORMAT

If you decide to resubmit your paper, please ensure that your manuscript complies with our editorial policies and complete and upload the checklist below as a Related Manuscript file type with the revised article:

Editorial Policy Policy requirements
(Download the link to your computer as a PDF.)

For Manuscripts that fall into the following fields:

- Behavioural and social science
- Ecological, evolutionary & environmental sciences
- Life sciences

An updated and completed version of our Reporting Summary must be uploaded with the revised manuscript

You can download the form here:

<https://www.nature.com/documents/nr-reporting-summary.zip>

For your information, you can find some guidance regarding format requirements summarized on the following checklist: (<https://www.nature.com/documents/commsj-phys-style-formatting-checklist-article.pdf>) and formatting guide (<https://www.nature.com/documents/commsj-phys-style-formatting-guide-accept.pdf>).

REVIEWER COMMENTS:

Reviewer #1 (Remarks to the Author):

This paper develops a data-driven seasonal forecast model for European heatwaves using machine learning. The proposed model has advantages in terms of both cost and skill compared to dynamical models. The authors demonstrate that their method matches or outperforms state-of-the-art dynamical seasonal forecast models from the Copernicus Climate Change Service (C3S) in certain regions, particularly over Scandinavia. However, the current methodology has room for improvement regarding methodological justification, validation, and a more detailed explanation of the methods. Detailed comments are provided below.

1. Page 2, line 49: The authors criticize previous studies for requiring pre-selection of predictor variables. However, this study also involves a feature selection process, which suggests that it may not be entirely novel in this aspect. Could you provide further discussion on how your approach differs from previous studies in predictor selection?
2. This paper employs the k-means clustering method for feature selection and extracts five clusters. How was the value of 'k' determined, and why was this specific number chosen? A more detailed justification is needed.
3. The paper uses various machine learning models, including Random Forest, Light Gradient Boost, AdaBoost, and Multi-Layer Perceptron. The performance of these models is known to be sensitive to parameter tuning, particularly for Gradient Boost, AdaBoost, and Perceptron models. However, the paper does not explain how these parameters were tuned or which parameters were chosen for model building. A detailed explanation is necessary to allow readers to follow and replicate the methodology.
4. The target variable is ERA5-based reanalysis maximum air temperature. However, I suggest including a section that validates or compares the model results with in-situ station-based temperature data in Europe. Given that high-quality in-situ station datasets are available in Europe, incorporating this additional validation and comparing data-driven and dynamical models would significantly strengthen the paper's robustness.
5. In the Introduction, the authors highlight extreme heatwave years such as 2003, 2010, and 2023. Could you add further discussion or analysis focusing on these years? Accurately estimating extreme heatwave years is crucial, as previous studies have struggled with this challenge.
6. Why not explore deep learning-based models such as LSTM or U-Net? Image-to-image prediction models have recently become more dominant and attractive due to their improved skill scores in weather prediction tasks. Could you add a deep learning model to your methodology or at least discuss its potential applicability?
7. While the study identifies key predictors (e.g., soil moisture, geopotential height, SST, etc.), further discussion is needed

on: Why certain predictors are dominant in specific regions.; The plausibility of the identified remote predictors (e.g., tropical Atlantic OLR) influencing European heatwaves.' Whether the identified predictors align with known physical mechanisms in climate science.

8. Please include a discussion on SHAP (Shapley Additive Explanations) values to assess the impact of key parameters. This would help explain the significance of newly identified key predictors, which is one of the major contributions of the paper.

Reviewer #2 (Remarks to the Author):

Review of "Feature selection for data-driven seasonal forecasts of European heatwaves" paper by McAdam et al.

The authors propose a novel data-driven forecasting system for seasonal forecasts of heatwaves using an evolutionary feature selection algorithm. The novelty of their approach consists of using a long simulation with a climate model for feature selection. This then allows to fit data-driven models relatively cheaply and may also provide some insight into drivers of heatwaves. They demonstrate that their data-driven approach is comparable in skill to dynamical systems, while being much cheaper to compute. The study area and proposed solution are relevant and novel and the manuscript is well written. Some of the choices of the paper, however, seem flawed as outlined below. Therefore, I recommend to accept the manuscript with major revisions.

Setup of the study

The setup for efficient modelling in low-predictability contexts such as seasonal forecasts is clearly challenging. I agree with the authors that there is room for alternative approaches to the dominant 'emulate dynamical models' paradigm that has gained heavy traction over the past years. And using relatively simple models to directly predict the quantity of interest (e.g. some characteristic of weather months ahead) surely has its value. There are two interesting aspects in this manuscript, namely the feature selection to aid in generalizing (through controlling overfitting) and understanding data-driven predictions and the use of a century-scale simulation to train data-driven seasonal forecasting. The benefit of both novelties is in my view not sufficiently documented.

To demonstrate that using the past2k to train the data-driven model one would need at least an alternative data-driven prediction based on a much shorter dataset (e.g. ERA5 pre 1993). To illustrate that the feature selection is instrumental, one would need at least a comparison with a prediction that uses all available features.

Focus on heatwaves

The authors ultimately use MSE for the feature selection and evaluate performance of the forecasts from their data-driven approach and benchmark dynamical system using correlation. While this is understandable and defensible for a general-purpose application, this choice seems somewhat at odds with their stated goal of assessing performance for heatwave forecasts. For such an application, I would expect a risk-based approach and probabilistic forecasting to be way more valuable for real-world applications.

The choice of loss function(MSE) also leads to noticeable regression-to-the-mean effects. The consequence is that the prediction is smoothed out, i.e. has too little inter-annual variability in this univariate time-series application. Alternative loss functions have been proposed to somewhat mitigate this effect in particular for applications for which arguably the high-intensity events (heatwaves) are the most relevant to forecast correctly. Having said that, it is obvious that only the move to a risk-based framing with probabilistic predictions allows to address the fundamental problem.

I am fully aware that this is a somewhat sweeping comment that can't be easily addressed without much further work. Therefore, I suggest the authors at least discuss the shortcomings of the current approach and how these could be overcome. One option that would be similar in complexity, but inherently probabilistic, would be to forecast threshold exceedance probabilities with a logistic regression model. Obviously, the evaluation would need to be adjusted accordingly.

Model selection

The authors fail to motivate why using an evolutionary approach for feature selection is necessary. Arguably, the same result could be achieved with well-established model selection approaches for linear models that are computationally way cheaper (e.g. forward/backward stepwise selection). As there is a plethora of algorithms for model selection available, I encourage the authors to better motivate their choice.

Minor comments

L89: the first implication is trivial and could be avoided. For the second implication, please elaborate how the similarity of the skill pattern with dynamical models demonstrates the limitations of the approach.

Figure 1: Why are there random under-performing solutions after 1000s of optimization steps far removed from the bulk of

solutions that show a very predictable development (i.e. the orange red dots in the right part of the relative point clouds)?

Figure 2 suggests that the method suffers from overfitting to the training data, as the performance on the test set is considerably worse compared with the training set. Is there any control of optimization to stop before such overfitting entails? If not, I suggest to split the training period in a training and validation period and use the latter to stop the optimization when overfitting occurs (i.e. when training and validation scores markedly start to differ).

Figure 2: I would love to see the same metric also for the ERA5 based test set. Not least to get an impression of the differences between past2k and ERA5.

Figure 4/5: Field significance of results: clearly there is no indication of significant differences. Under the null hypothesis (and iid assumptions), you would expect at least 5% of the grid boxes to show significant differences. With correlation between grid boxes, you could even get a much larger fraction under the null.

L224: typesetting error in units

L246: It is not clear to me how NDQ90 for the dynamical systems and the past2k simulation has been computed. Are values normalized using quantiles estimated from ERA5 or from their own 'climatology'?

L294: I assume the different data-driven approaches that are ultimately applied to ERA5 are also fitted on the training set of past2k (0-1600). Could you please clarify and state the setup more clearly?

L328: In the interest of reproducibility, I encourage the authors to make the pre-processed features and targets available as an analysis-ready dataset.

Reviewer #3 (Remarks to the Author):

In this manuscript, the authors used the optimization-based feature selection method to select different climate variables and constructed a prediction model based on regression and machine learning methods, focusing on the seasonal prediction of European heatwaves. The manuscript has reasonable logic and clear expression; however, I have the following major concerns that require revision.

Abstract needs to be improved, for example, skill improvement and what the key indicators are should be clearly stated.

Line 261-262: The clustering method with $k=5$ may not be able to completely separate the modes of different regions and different variables. Using Varimax-EOF for modal decomposition may be more helpful for feature extraction. If possible, the author can consider comparing the performance with other dimensionality reduction methods.

Line 276-277: The mode integration results used by the author start from the 0th year. Did the author consider the spin-up problem? Are the mode integration results from the 0th to the 100th year stable? This will directly affect the quality of the model simulation.

Line 247-317: The author can add a schematic diagram of the model structure and experimental process to help readers understand MMEA.

Line 290-291: The author emphasized in the title that the main work of this article is feature selection, but the details of the feature selection part were not elaborated, such as what the optimization is. Is the loss function reduced by backpropagation, or is the hyperparameter adjusted randomly? The author should briefly introduce the optimization method in the article.

Ext Data Fig.2 'Cluster maps are shown in Fig. X.': Missing the specific image number.

Line 213: The author only selected the simulation results of one model, but did not evaluate the simulation of MPI-ESM paleo-simulation. It is recommended that the author supplement the evaluation results of the model to enhance the credibility of subsequent conclusions.

It is recommended that the author discuss the comparison between the feature selection method in this article and other commonly used feature selection methods (Pearson correlation coefficient method, distance correlation coefficient method, etc.) to reflect the advantages of this method.

As far as I understand, the major content of the manuscript focuses on seasonal predictions of European heat waves. Although one feature selection method is used, the author lacks comparison with other feature selection methods. Therefore, I think the title of this article can be modified accordingly unless the authors put more efforts on the feature selection method.

The prediction skills of the data-driven method in this paper are comparable to those of the dynamical prediction models. Does this mean that the feature selection method is not advantageous? Compared with the high prediction skills obtained by

previous end-to-end data-driven methods (Chattopadhyay et al., 2020, etc.), the author may want to add discussions on this.

Reference:

Chattopadhyay, A., Nabizadeh, E., & Hassanzadeh, P. (2020). Analog forecasting of extreme-causing weather patterns using deep learning. *Journal of Advances in Modeling Earth Systems*, 12(2), e2019MS001958.

** Visit Nature Research's author and referees' website at www.nature.com/authors for information about policies, services and author benefits**

Communications Earth & Environment is committed to improving transparency in authorship. As part of our efforts in this direction, we are now requesting that all authors identified as 'corresponding author' create and link their Open Researcher and Contributor Identifier (ORCID) with their account on the Manuscript Tracking System prior to acceptance. ORCID helps the scientific community achieve unambiguous attribution of all scholarly contributions. You can create and link your ORCID from the home page of the Manuscript Tracking System by clicking on 'Modify my Springer Nature account' and following the instructions in the link below. Please also inform all co-authors that they can add their ORCIDs to their accounts and that they must do so prior to acceptance.

<https://www.springernature.com/gp/researchers/orcid/orcid-for-nature-research>

For more information please visit <http://www.springernature.com/orcid>

If you experience problems in linking your ORCID, please contact the Platform Support Helpdesk.

Version 1:

Decision Letter:

** Please ensure you delete the link to your author home page in this e-mail if you wish to forward it to your coauthors **

Dear Dr McAdam,

Your manuscript titled "Feature selection for data-driven seasonal forecasts of European heatwaves" has now been seen by our reviewers, whose comments appear below. In light of their advice we are delighted to say that we are happy, in principle, to publish a suitably revised version in *Communications Earth & Environment*.

We therefore invite you to edit your manuscript to comply with our format requirements and to maximise the accessibility and therefore the impact of your work.

EDITORIAL REQUESTS:

Please review our specific editorial comments and requests regarding your manuscript in the attached "Editorial Requests Table".

*****Please take care to match our formatting and policy requirements. We will check revised manuscript and return manuscripts that do not comply. Such requests will lead to delays. *****

Please outline your response to each request in the right hand column. Please upload the completed table with your manuscript files as a Related Manuscript file.

If you have any questions or concerns about any of our requests, please do not hesitate to contact me.

SUBMISSION INFORMATION:

In order to accept your paper, we require the files listed at the end of the Editorial Requests Table; the list of required files is also available at <https://www.nature.com/documents/commsj-file-checklist.pdf>.

OPEN ACCESS:

Communications Earth & Environment is a fully open access journal. Articles are made freely accessible on publication. For further information about article processing charges, open access funding, and advice and support from Nature Portfolio, please visit <https://www.nature.com/commsenv/open-access>

At acceptance, you will be provided with instructions for completing the open access licence agreement on behalf of all authors. This grants us the necessary permissions to publish your paper. Additionally, you will be asked to declare that all required third party permissions have been obtained, and to provide billing information in order to pay the article-processing charge (APC).

Please use the following link to submit the above items:

Link Redacted

** This url links to your confidential home page and associated information about manuscripts you may have submitted or be reviewing for us. If you wish to forward this email to co-authors, please delete the link to your homepage first **

We hope to hear from you within two weeks; please let us know if you need more time.

Best regards,

Seung-Ki Min, PhD
Editorial Board Member
Communications Earth & Environment

Martina Grecequet, PhD
Senior Editor,
Communications Earth & Environment
Consulting Editor
Communications Sustainability

REVIEWERS' COMMENTS:

Reviewer #1 (Remarks to the Author):

This paper has been seriously revised by the authors, and they have answered all the questions which I concerned about.

Reviewer #2 (Remarks to the Author):

Second review of "Feature selection for data-driven seasonal forecasts of European heatwaves" by McAdam et al.

The authors have revised the manuscript considerably and have sufficiently addressed all open comments. Thank you for the thorough revision. The manuscript is now ready for publication.

Reviewer #3 (Remarks to the Author):

The revised manuscript has addressed my concerns, and therefore I am happy to recommend acceptance.

** Visit Nature Portfolio's author and referees' website at www.nature.com/authors for information about policies, services and author benefits**

Open Access This Peer Review File is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made.

In cases where reviewers are anonymous, credit should be given to 'Anonymous Referee' and the source.

The images or other third party material in this Peer Review File are included in the article's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder.

To view a copy of this license, visit <https://creativecommons.org/licenses/by/4.0/>

We thank the reviewers for their constructive comments and suggestions which we believe have contributed to improving the clarity of the manuscript and help display the robustness of results.

Based on the comments of all reviewers, we have performed a substantial amount of extra tests, including:

- Increasing the number of clusters used to define the predictors to $k=10$.
- Including a SHapley Additive exPlanations (SHAP) analysis to further explore the most-frequently selected features.
- Evaluating the temperature and heatwave statistics in the paleo-simulation used for model training.
- Comparing to a cheaper feature selection method, based on correlation of the predictors to the target.

These tests served to increase the robustness of results, and for this we are grateful to the reviewers. Ultimately, and encouragingly, the main results of the paper have not changed but are further supported by the additions.

We have added several new figures to the manuscript:

- **Figure X: Forecasts of summer heatwaves over northern Italy**
- **Ext. Data Figure 1: Schematic of the optimisation-based feature selection framework.**
- **Ext. Data Figure 6: Comparison of temperature and heatwave statistics in ERA5 and past2k.**
- **Ext. Data Figure 7: Geographical patterns of selected features and SHAP analysis.**
- **Ext. Data Figure 10: Maps of large-scale heatwaves over Europe in 2003 and 2010.**
- **Ext. Data Figure 11: Skill score (correlation) of forecasts derived from correlation-based feature selection method.**

For the sake of not lengthening the paper excessively, we have removed the previous Figure 6 (Correlation skill-score using only local or non-local skill). The main conclusions from that figure can be drawn from other Figures (e.g. current Figure 3).

Our responses are written in bold font, and figures appearing in the responses are referred to as “Plots”.

Reviewer #1 (Remarks to the Author):

This paper develops a data-driven seasonal forecast model for European heatwaves using machine learning. The proposed model has advantages in terms of both cost and skill compared to dynamical models. The authors demonstrate that their method matches or outperforms state-of-the-art dynamical seasonal forecast models from the Copernicus Climate

Change Service (C3S) in certain regions, particularly over Scandinavia. However, the current methodology has room for improvement regarding methodological justification, validation, and a more detailed explanation of the methods. Detailed comments are provided below.

1. Page 2, line 49: The authors criticize previous studies for requiring pre-selection of predictor variables. However, this study also involves a feature selection process, which suggests that it may not be entirely novel in this aspect. Could you provide further discussion on how your approach differs from previous studies in predictor selection?

We thank the reviewer for highlighting the lack of clarity in this key point. We now specify that previous papers use a limited number of predictors, typically based on known phenomena.

Lines 40-51: “Prior to the widespread use of ML, statistical seasonal forecasting showed that the input of known predictors, such as soil moisture for European heatwaves, into simple statistical models could provide skill for the prediction of certain climate variables on seasonal timescales [31, 32]. Previously, such methods relied on the selection of known drivers and thus were limited by the current scientific understanding of heatwave dynamics. Nowadays, more sophisticated ML models are used and are able to select from a set of potential predictors, as widely demonstrated on the subseasonal timescale [33, 34, 35, 36, 37]. ML-based seasonal forecasting techniques have shown that spatially and temporally distant predictors of seasonal climate can be identified and used to make accurate predictions [38, 39, 40, 41, 42]. However, studies either do not include comparisons to operational systems or use a restricted number of pre-selected and known predictors in their feature selection. Moreover, there is currently no purely data-driven seasonal forecast approach for Europe, nor one that focuses explicitly on temperature extremes”.

Furthermore, in the discussion, we comment on the novelty of MMEA and the potential of using a wider pool of predictors.

Lines 201-217: “Here, the pool of potential predictors used is wider than typically used in feature selection studies on S2S or seasonal forecasting; the number of cluster variables is 70, with each time-lag (up to 28 weeks prior to the target) counting as an individual predictor thus leading to a total of roughly 2000 potential predictors. The framework does not rely only on expected or known predictors, but instead clusters predictor variables as a means of reducing the dimensionality of the problem without including human bias or relying on prior knowledge. A benefit of the optimisation algorithm used is its capacity to filter out unnecessary predictor information and identify the key predictors, thereby allowing the input of many potential predictors. Across the domain studied here, between 3.2 and 10.6% (6.5% on average) of the cluster-lag combinations are selected as predictors. Generally, including more predictors slows down ML forecasts, ruling out the possibility of including all 2000 in this case. To illustrate the benefits of the optimisation-based method, we compare it to an alternative and simpler method of feature selection

based on linear correlation analysis, in which the selected predictors are those significantly (positively or negatively) correlated to the target data during the period 1993-2016. Although the correlation approach selects a similar number of predictors (7.8% on average across the domain), the resulting forecasts are considerably poorer in skill (Ext. Data Fig. 11). This highlights the greater ability of our optimisation-based approach to capture physically plausible predictors compared to simpler statistical-based approaches”.

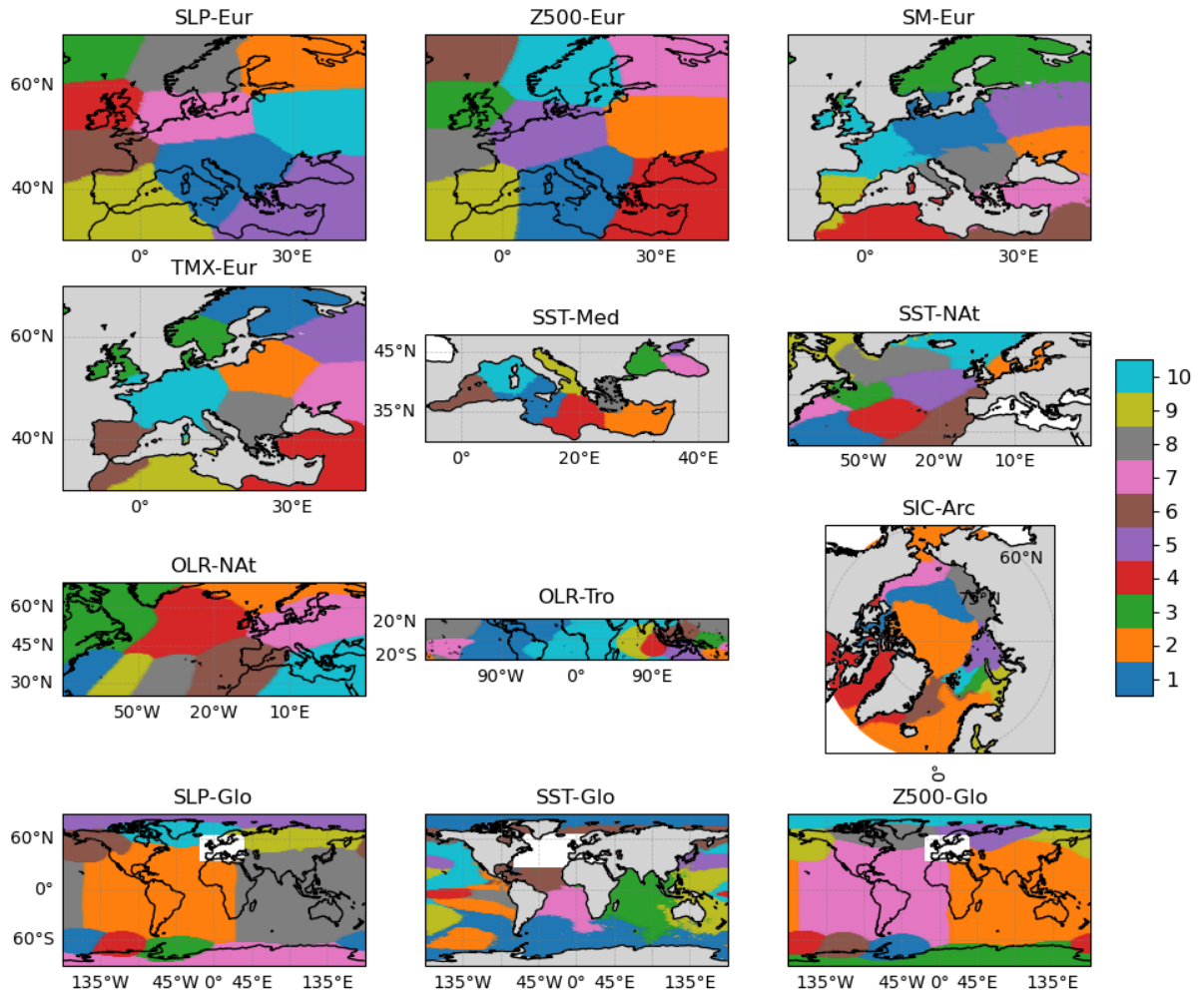
2. This paper employs the k-means clustering method for feature selection and extracts five clusters. How was the value of ‘k’ determined, and why was this specific number chosen? A more detailed justification is needed.

We thank the reviewer for raising this point. $k=5$ is an arbitrary choice made to demonstrate the capability of the framework with large dimensionality reduction. The framework permits the use of any potential predictors the user wishes to study. We add the following line to the Methods:

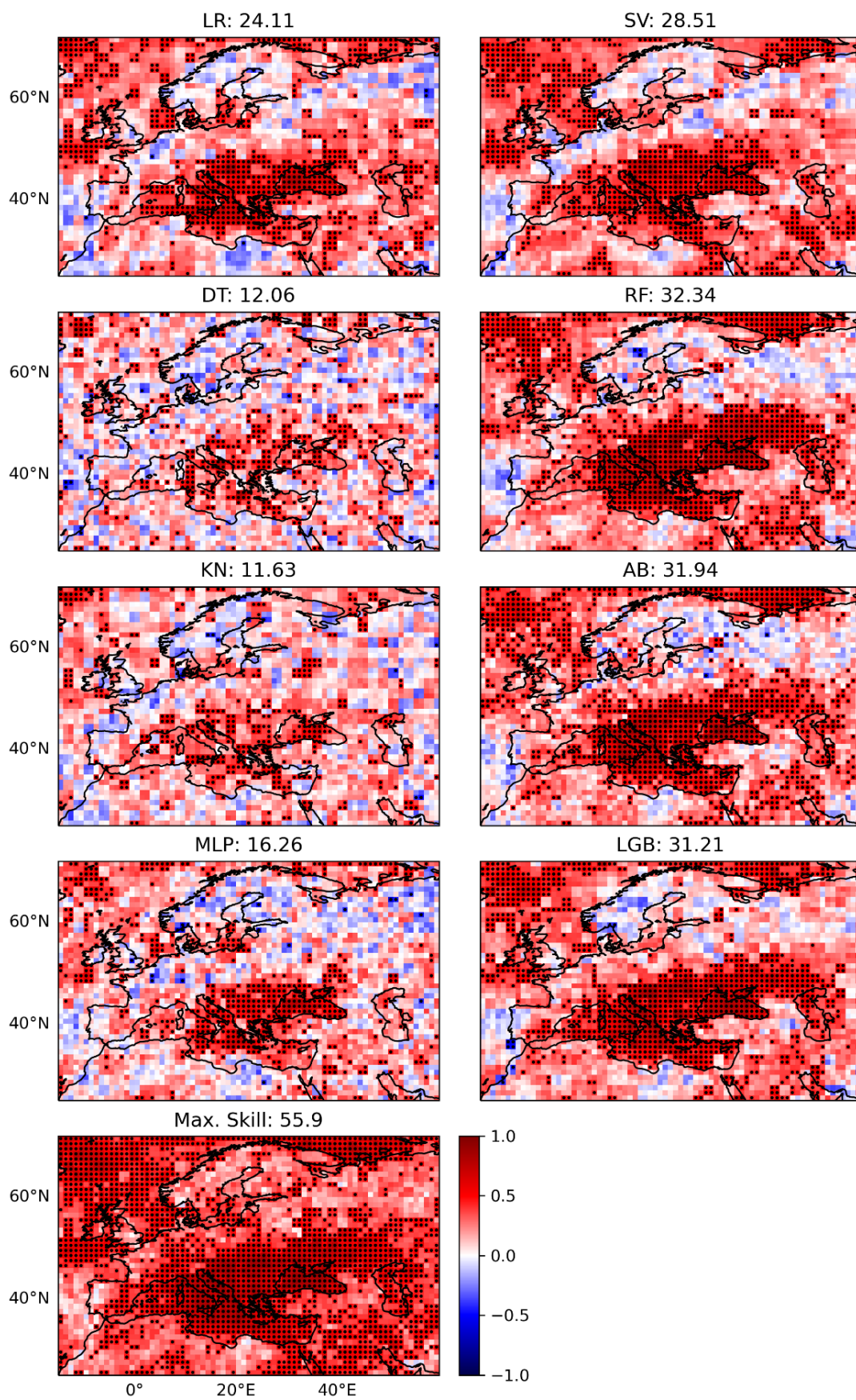
Lines 320-323: *“Given that our objective is to demonstrate the capability of such a framework, we use $k=5$ as a compromise between choosing a large pool of predictors and performing a suitable reduction of the dimensionality of the problem.”*

To test the sensitivity of results to the choice of potential predictors, we have now also performed an equivalent experiment using $k=10$ for each variable-domain (i.e. doubling the number of predictors; Plot 1.1). The optimisation took roughly 2-3 as many CPU hours as for the case in which $k=5$. The resulting data-driven seasonal forecasts are shown in Plot 1.2, and display negligible differences in skill to the experiment with $k=5$ clusters (Ext. Data Fig. 8). Thus, there is apparently little sensitivity to the cluster number.

The lack of added skill when using more potential predictors implies that the large-scale information gained from training/optimising on the paleo simulation is sufficiently represented by $k=5$, and that there is no clear added-value provided by choosing $k=10$. Performing an extensive analysis on dimensionality reduction is not the scope of this paper, and thus we choose not to include this analysis in the paper.



Plot 1.1. Clusters of variables used to construct time series of potential predictors. Variable labels: mean sea level pressure (SLP), geopotential height at 500hPa (z500), soil moisture (SM), daily maximum 2m temperature (TMX), sea surface temperature (SST), outgoing longwave radiation (OLR), and sea ice concentration (SIC). K-means clustering is applied to the daily ERA5 time series of anomalies for the period 1981-2010, with $k=10$. White indicates zones excluded from clustering, while grey indicates land (e.g. SST) or the ocean (e.g. for SM).



Plot 1.2. Skill score (correlation) of forecasts of May-July NDQ90 in the data-driven approach using a range ML models over the forecast test period 1993-2016, when using increased number of clusters (k=10) as predictors. The following regression ML models are used: Linear Regression (LR), Support Vector (SV), Decision Tree (DT), Random Forest (RF), K-nearest neighbour (KN), AdaBoost (AB), Multi-Layer Perceptron (MLP), Light Gradient Boost (LGB). The skill corresponding to the best model in each grid point, used for the data-driven approach (Fig. 5) is shown by Max. Skill. Black stippling represents statistically significant correlation at the 95% confidence interval. Numbers in the titles correspond to the percentage of the domain covered by significant skill.

3. The paper uses various machine learning models, including Random Forest, Light Gradient Boost, AdaBoost, and Multi-Layer Perceptron. The performance of these models is known to be sensitive to parameter tuning, particularly for Gradient Boost, AdaBoost, and Perceptron models. However, the paper does not explain how these parameters were tuned or which parameters were chosen for model building. A detailed explanation is necessary to allow readers to follow and replicate the methodology.

We thank the reviewer for identifying this missing point. Here, we use the default parameters of the ML model modules (e.g. SciKit), which themselves are based on extensive research and represent reasonable starting points that work well across diverse datasets, and represent a compromise between computational expense and complexity. We have added this point to the Methods and Code Availability statement, plus the versions of the modules used which is sufficient for readers to track down the default parameters.

Lines 358-361: “While the optimisation is based on a linear regression forecast for the sake of computational time, the second step of producing real-world forecasts can be performed with ML models in order to boost skill. Several candidates are tested (Ext. Data. Fig. 8), each using the default values provided in the accessible Python modules (see Code Availability statement).”

Lines 399-406: “The machine learning models used in this study are from the scikit-learn (v1.4.1) python programming language package, and Microsoft lightgbm (v4.3.0). The default parameter values corresponding to the specific versions are used.”

A greater understanding and quantification of ML model set-up effects, on this problem and in climate applications in general, is beyond the scope of this study. Performing a parameter tuning for each grid point in each model would be an enormous undertaking. We fully agree this needs to be considered in future developments of feature selections and forecasting, and add this to the discussion:

Lines 229-231: “Although already successful, future improvements can be made to this prediction system. Parameter tuning for the ML models used could identify potential improvements, but it is an enormous undertaking for several models covering a wide geographical domain.”

4. The target variable is ERA5-based reanalysis maximum air temperature. However, I suggest including a section that validates or compares the model results with in-situ station-based temperature data in Europe. Given that high-quality in-situ station datasets are available in Europe, incorporating this additional validation and comparing data-driven and dynamical models would significantly strengthen the paper's robustness.

We thank the reviewer for this suggestion, and respectfully disagree with the necessity of adding in-situ data to this study. In the S2S forecasting communities, ERA5 is a common validation dataset for regional-scale patterns - this is particularly true for seasonal forecast validation of temperature (e.g. Johnson et al., 2019) and heatwaves (Prodhomme et al., 2023; Torralba et al., 2024). Moreover, previous studies have demonstrated the capability of ERA5 in representing temperature and its extremes (Velikuo et al., 2022; Skrynyk et al., 2025).

The reviewer nonetheless is right to bring up potential biases and differences between ERA5 and in-situ observations, and so we add the following to the description of ERA5 in the Methods section.

Lines 250-252: *"Previous studies have shown that ERA5 accurately reproduces both mean and extreme temperatures [58], confirming that it is a reliable source of climate information over Europe, in particular for heatwave indicators [59]"*.

[58] Velikou, K., Lazoglou, G., Tolika, K. & Anagnostopoulou, C. Reliability of the era5 in replicating mean and extreme temperatures across europe. *Water* 14, 543 (2022).

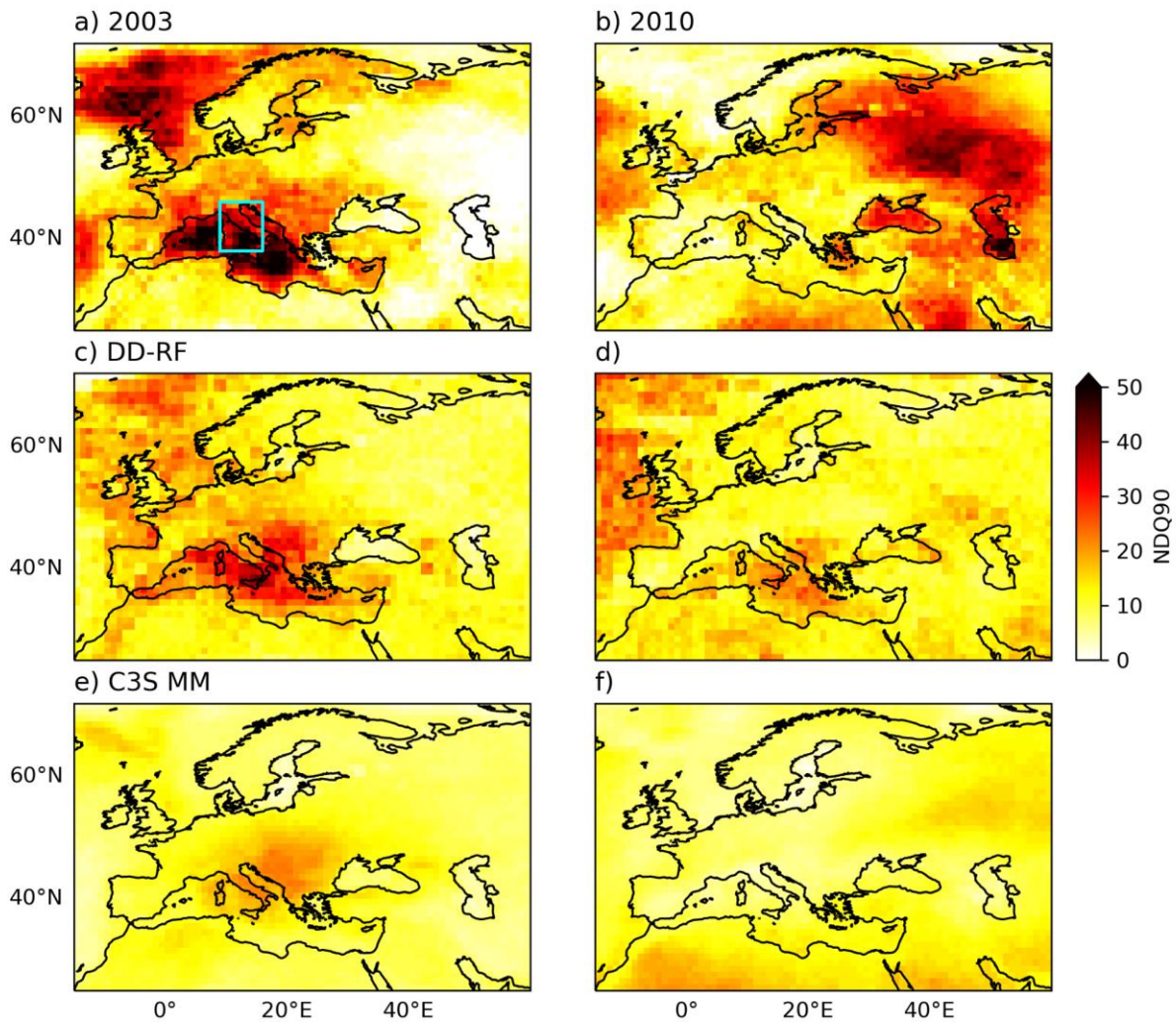
[59] Skrynyk, O., Aguilar, E. & Cimolai, C. The sensitivity of heatwave climatology to input gridded datasets: A case study of ukraine. *Atmosphere* 16, 289 (2025).

5. In the Introduction, the authors highlight extreme heatwave years such as 2003, 2010, and 2023. Could you add further discussion or analysis focusing on these years? Accurately estimating extreme heatwave years is crucial, as previous studies have struggled with this challenge.

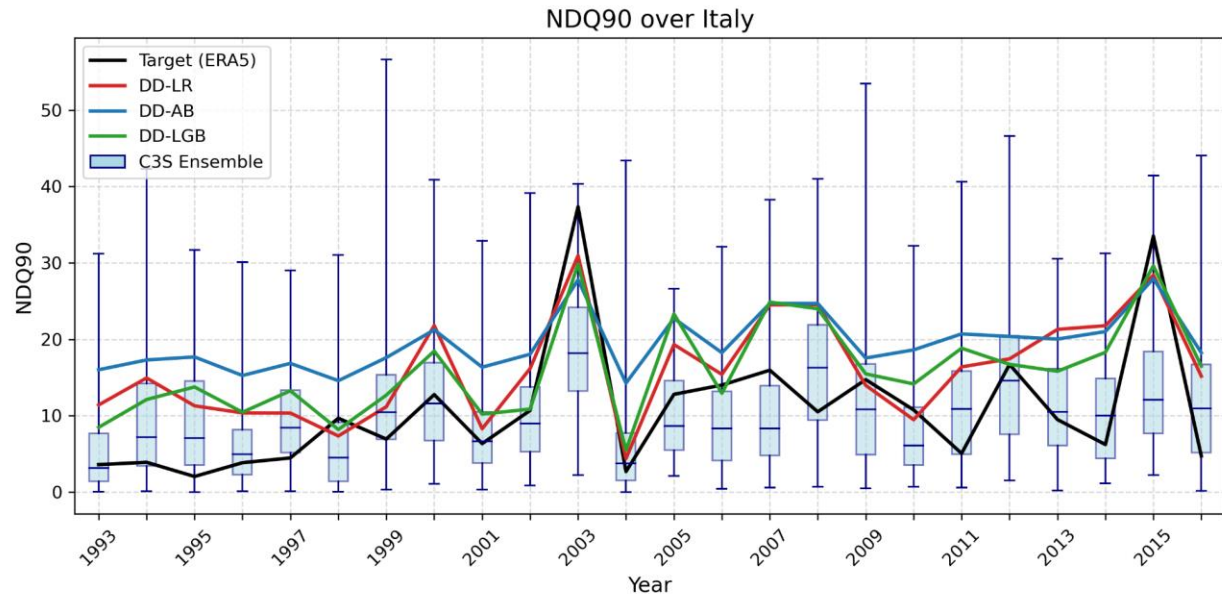
We thank the reviewer for this suggestion, which we agree with completely. We now include maps of specific summers (2003 and 2010; Ext. Data Fig. 10/Plot 1.3), and a time series of forecasts for northern Italy (Figure 5/Plot 1.4).

Lines 158-167: *"It is crucial for newly proposed systems to demonstrate skill in forecasting the most exceptional events and, crucially, the data-driven forecasts display this capability in some cases (Fig. 6). In northern Italy, where both data-driven and dynamical systems generally display high skill, the top-performing models in data-driven system forecasts are remarkably close to the observed values for the two years with the greatest number of HW days (2003 and 2015). In this region, a simple linear regression model is as effective as ML-based models and outperforms the dynamical systems (Ext. Data Fig. 8), while some display stronger biases in NDQ90 (e.g. Light Gradient Boost -*

DD-LGB) than others. The extent of the HW in 2003 across western Europe and the Mediterranean basin is also well forecast by the data-driven approach (Ext. Data Fig. 10), although the exceptionally deadly event of 2010 over western Russia was not predicted by either type of forecast [20]”.



Plot 1.3: Maps of large-scale heatwaves over Europe in 2003 and 2010. Number of HW days (NDQ90) from May to July in ERA5 (a & b), the data-driven approach with a Random Forest model (c & d) and the C3S multi-model median (e & f). The green box denotes the Northern Italy area used in Fig. 6.



Plot 1.4: Forecasts of NDQ90 summer heatwave indicators over northern Italy. ERA5 (black) is compared to the full ensemble (120 members) of the dynamical systems (C3S Ensemble) and the data-driven approach with three high-performing ML models (Linear Regression, DD-LR; AdaBoost, DD-AB; Light Gradient Boost, DD-LGB). Boxplots represent the medians, interquartile ranges and maxima and minima for each forecast year in the C3S Ensemble. The correlation skill scores over the 1993-2016 period are as follows: C3S ensemble median (0.63), DD-LR (0.74), DD-AB (0.78) and DD-LGB (0.76). The box used to define northern Italy is shown in Ext. Data Fig. 10.

6. Why not explore deep learning-based models such as LSTM or U-Net? Image-to-image prediction models have recently become more dominant and attractive due to their improved skill scores in weather prediction tasks. Could you add a deep learning model to your methodology or at least discuss its potential applicability?

We note that Multi-Layer Perceptron - a feedforward neural network - is already used in this study and does not outperform some ML models like Random Forests in our analysis. We have now added this to the analysis of candidate models.

Lines 367-370: “Multi-Layer Perceptron is a deep learning (DL) neural-network model, ideal for larger datasets in which there are more non-linear relationships between predictors and target. In this study, it is outperformed by ML-based models such as Random Forest, which has been found to be more suited to similar tasks [34, 30].”

The use of deep learning in this context, while indeed possible, in our opinion represents a different approach. Deep learning is an end-to-end approach, and is not typically preceded by feature selection. Our framework instead implements inherent explainability by allowing for dimensionality reduction and selection of ML models. The application of

deep learning to the problem of seasonal forecasting is indeed interesting but a comparison between the two approaches is beyond the scope of this paper.

7. While the study identifies key predictors (e.g., soil moisture, geopotential height, SST, etc.), further discussion is needed on: Why certain predictors are dominant in specific regions.; The plausibility of the identified remote predictors (e.g., tropical Atlantic OLR) influencing European heatwaves.’ Whether the identified predictors align with known physical mechanisms in climate science.

Answer included with the next point, see below.

8. Please include a discussion on SHAP (Shapley Additive Explanations) values to assess the impact of key parameters. This would help explain the significance of newly identified key predictors, which is one of the major contributions of the paper.

We thank the reviewer for this suggestion, which complements and expands our analysis on predictors. Here, we respond to points 7 & 8.

We have added the geographical distribution of key selected features in Ext. Data Figure 7 (Plot 1.5). First, we highlight the grid points in which the four most-commonly selected features are selected in past2k (left column). Then, we employ the SHapley Additive exPlanations (SHAP) analysis to predictions of 1993-2016 with Random Forests, to quantify the contribution of each target predictor to the forecast. SHAP confirms and quantifies the role of key predictors e.g that local nearby Tmax provides more predictability than, for example, SST in the Tropical Pacific.

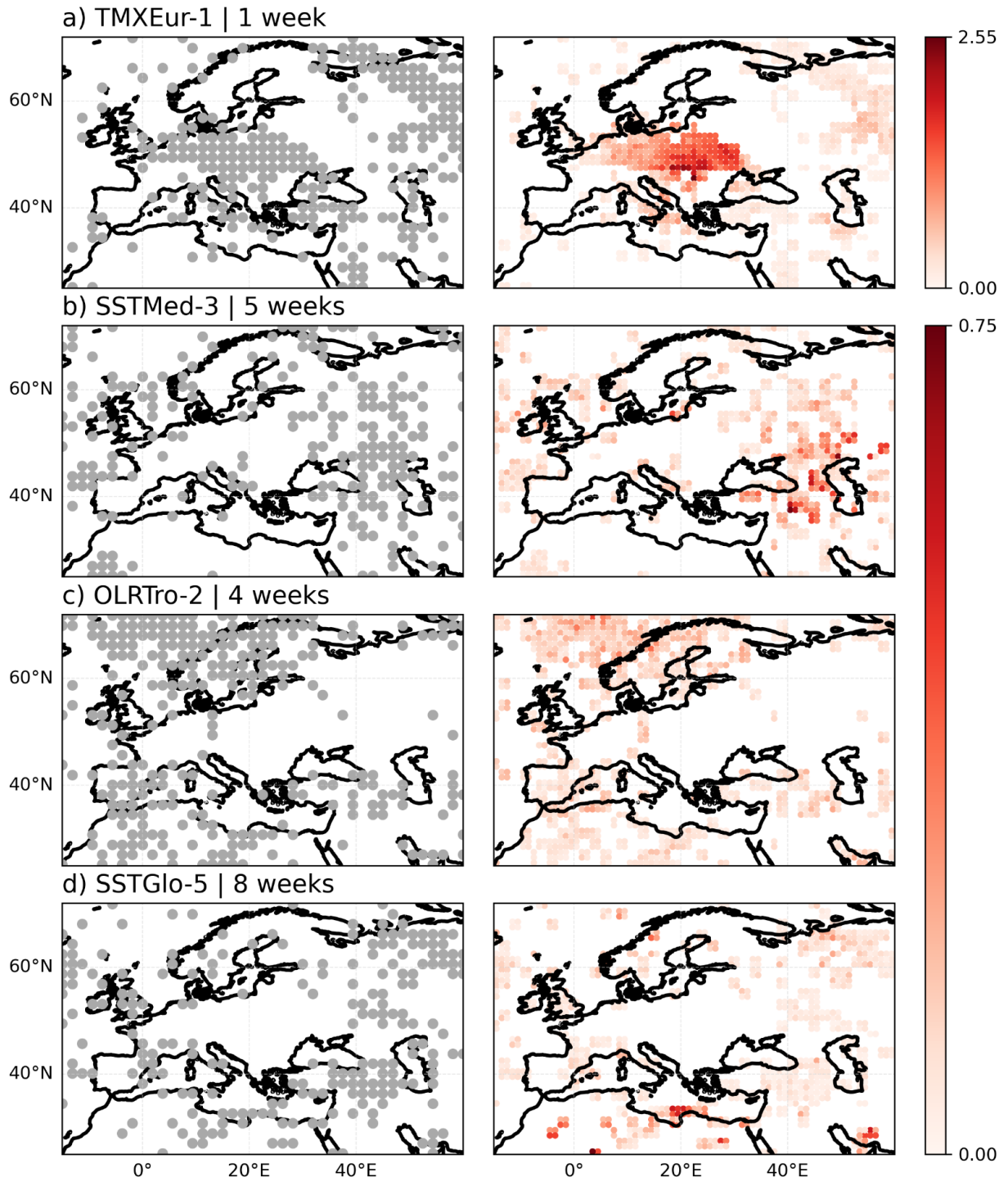
The following has been added to the Results section.

Lines 99-126: “Collecting the optimal predictors from all individual points across Europe (see Ext. Data Fig. 3 for an example) provides an overview of the model-world HW drivers at a regional level (Fig. 3). The most commonly selected variables across the domain are the European soil moisture, temperature and geopotential height (z500) clusters. The identified key role of these local predictors agrees with studies of many HWs that have occurred in Europe [31, 15]. Commonly selected predictors that represent more distant precursors include sea surface temperature (SST) over the equatorial Pacific and outgoing longwave radiation (OLR) over the tropical Atlantic. While the former represents the phase of the El Niño Southern Oscillation, which is known to play a role in European climate extremes [47], the contribution of the OLR over the tropical Atlantic is not obvious. The feature selection also allows us to study the time lags in which the variables play a role. The most frequently selected time-lags occur on average around six weeks prior to initialisation (i.e. mid-March; Fig. 3, Ext. Data Fig. 6). However, the key temporal lag depends on the variables. Temperature and z500 clusters are selected more frequently in the few weeks prior to initialisation and decay gradually with longer lag, while soil moisture and sea ice selection peak between 7-8 weeks prior to initialisation. The number of predictors selected before February is negligible. The most commonly

selected short-term European predictors are unsurprisingly selected within or near to the area they represent (Ext. Data Fig. 7); TMXEur-1 at 1 week time-lag is selected across central Europe while SSTMed-3 at 5 weeks time-lag is largely selected around the Black Sea. For the geographically distant predictors, the tropical Atlantic OLR (OLRTro-2 at 4 weeks lag) is chosen as a predictor for areas across the Barents Sea and Scandinavia, as well as parts of the southern Mediterranean Sea, while the selection of the tropical Pacific SST is more sporadic. Studying these points in a SHAP (SHapley Additive exPlanations [48]; Methods) analysis, used to quantify their relative contribution to forecasts, confirms that the local predictors carry more predictive value, while the more distant predictors contribute more weakly (Ext. Data Fig. 7). While distant teleconnection-based predictors would be expected to have less direct impact on HW occurrence and therefore a relatively low predictive power according to the SHAP analysis, we cannot rule out that such drivers are present only in the model-world. However, their widespread selection, in particular that of OLR, suggests they are not spurious results and merit further analysis in future studies.”

While a description of SHAP has been added to the Methods.

Lines 383-393: *“SHAP is a method used to interpret machine learning models by quantifying the contribution of each feature/predictor to individual predictions. Here, we apply SHAP to Random Forest forecasts to explore the contribution of predictors selected in past2k to the forecasts using ERA5 predictors from 1993-2016. For each example predictor studied (Ext. Data Fig. 7), and in each grid cell, we calculate the average of the SHAP value magnitudes for the target predictors”.*



Plot 1.5. Geographical analysis of most selected predictors and time-lag. The locations in which the most frequently selected features from the past2k optimisation across Europe (as identified in Fig. 3) are selected (Left column). The absolute average SHAP values when used in the Random Forest model (Right column); higher values display more contribution of predictor to forecasts.

Note the reduced range for b-d. The percentages of the total domain in which the features are selected are 24.8% (a), 20.5 (b), 26.6 (c) and 20.5 (d).

Reviewer #2 (Remarks to the Author):

Review of “Feature selection for data-driven seasonal forecasts of European heatwaves” paper by McAdam et al.

The authors propose a novel data-driven forecasting system for seasonal forecasts of heatwaves using an evolutionary feature selection algorithm. The novelty of their approach consists of using a long simulation with a climate model for feature selection. This then allows to fit data-driven models relatively cheaply and may also provide some insight into drivers of heatwaves. They demonstrate that their data-driven approach is comparable in skill to dynamical systems, while being much cheaper to compute. The study area and proposed solution are relevant and novel and the manuscript is well written. Some of the choices of the paper, however, seem flawed as outlined below. Therefore, I recommend to accept the manuscript with major revisions.

Setup of the study

The setup for efficient modelling in low-predictability contexts such as seasonal forecasts is clearly challenging. I agree with the authors that there is room for alternative approaches to the dominant ‘emulate dynamical models’ paradigm that has gained heavy traction over the past years. And using relatively simple models to directly predict the quantity of interest (e.g. some characteristic of weather months ahead) surely has its value. There are two interesting aspects in this manuscript, namely the feature selection to aid in generalizing (through controlling overfitting) and understanding data-driven predictions and the use of a century-scale simulation to train data-driven seasonal forecasting. The benefit of both novelties is in my view not sufficiently documented.

We thank the reviewer for their positive feedback, and for pointing out that more explicit mention of the novelty is needed. This has been addressed in the introduction and the discussion.

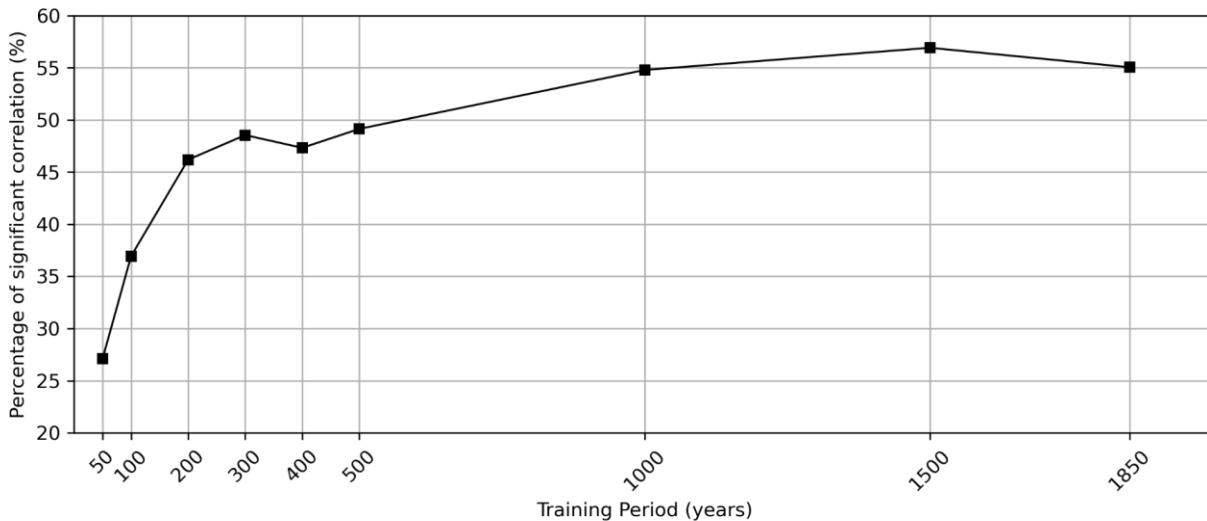
Lines 40-51: “Prior to the widespread use of ML, statistical seasonal forecasting showed that the input of known predictors, such as soil moisture for European heatwaves, into simple statistical models could provide skill for the prediction of certain climate variables on seasonal timescales [31, 32]. Previously, such methods relied on the selection of known drivers and thus were limited by the current scientific understanding of heatwave dynamics. Nowadays, more sophisticated ML models are used and are able to select

from a set of potential predictors, as widely demonstrated on the subseasonal timescale [33, 34, 35, 36, 37]. ML-based seasonal forecasting techniques have shown that spatially and temporally distant predictors of seasonal climate can be identified and used to make accurate predictions [38, 39, 40, 41, 42]. However, studies either do not include comparisons to operational systems or use a restricted number of pre-selected and known predictors in their feature selection. Moreover, there is currently no purely data-driven seasonal forecast approach for Europe, nor one that focuses explicitly on temperature extremes.”

Lines 201-217: “Here, the pool of potential predictors used is wider than typically used in feature selection studies on S2S or seasonal forecasting; the number of cluster variables is 70, with each time-lag (up to 28 weeks prior to the target) counting as an individual predictor thus leading to a total of roughly 2000 potential predictors. The framework does not rely only on expected or known predictors, but instead clusters predictor variables as a means of reducing the dimensionality of the problem without including human bias or relying on prior knowledge. A benefit of the optimisation algorithm used is its capacity to filter out unnecessary predictor information and identify the key predictors, thereby allowing the input of many potential predictors. Across the domain studied here, between 3.2 and 10.6% (6.5% on average) of the cluster-lag combinations are selected as predictors. Generally, including more predictors slows down ML forecasts, ruling out the possibility of including all 2000 in this case. To illustrate the benefits of the optimisation-based method, we compare it to an alternative and simpler method of feature selection based on linear correlation analysis, in which the selected predictors are those significantly (positively or negatively) correlated to the target data during the period 1993-2016. Although the correlation approach selects a similar number of predictors (7.8% on average across the domain), the resulting forecasts are considerably poorer in skill (Ext. Data Fig. 11). This highlights the greater ability of our optimisation-based approach to capture physically plausible predictors compared to simpler statistical-based approaches.”

To demonstrate that using the past2k to train the data-driven model one would need at least an alternative data-driven prediction based on a much shorter dataset (e.g. ERA5 pre 1993). To illustrate that the feature selection is instrumental, one would need at least a comparison with a prediction that uses all available features.

As we explore in the Extended Data Fig. 12 (Plot 2.1 below), using a training period of 50 years (of past2k) leads to only roughly 30% of the domain displaying a significant forecast skill. There is a sharp increase when moving from 50 years to 100s of years of data. Relying on ERA5 would mean using a 53-year training period (1940-1993).



Plot 2.1: Effect of training time on data-driven seasonal forecast skill. The y-axis represents the area of the European domain in which significant positive correlation is found. The training period represents the amount of years (starting from the year 0) on which the data-driven forecast model is trained.

Another key issue is that the HW index is defined relative to the 1993-2016 baseline period, as per the dynamical systems. This means there are many years of no HW days in the early half of the record, and would hinder any training.

Lastly, in response to the comment on using all available features, using all 2000 cluster-variable-lag combinations in a ML model, in particular Random Forests or Gradient Boosting, requires considerably more computational resources. In the example shown in Ext. Data Fig. 3, the feature selection extracts 4% of the total predictors. This strict selection is a major motivation for using the feature selection, shows that it is working effectively, and provides a significantly more efficient means of forecasting. We add to the discussion:

Lines 206-211: *“A benefit of the optimisation algorithm used is its capacity to filter out unnecessary predictor information and identify the key predictors, thereby allowing the input of many potential predictors. Across the domain studied here, between 3.2 and 10.6% (6.5% on average) of the cluster-lag combinations are selected as predictors. Generally, including more predictors slows down ML forecasts, ruling out the possibility of including all 2000 in this case.”*

Focus on heatwaves

The authors ultimately use MSE for the feature selection and evaluate performance of the forecasts from their data-driven approach and benchmark dynamical system using correlation. While this is understandable and defensible for a general-purpose application, this choice

seems somewhat at odds with their stated goal of assessing performance for heatwave forecasts. For such an application, I would expect a risk-based approach and probabilistic forecasting to be way more valuable for real-world applications.

The choice of loss function (MSE) also leads to noticeable regression-to-the-mean effects. The consequence is that the prediction is smoothed out, i.e. has too little inter-annual variability in this univariate time-series application. Alternative loss functions have been proposed to somewhat mitigate this effect in particular for applications for which arguably the high-intensity events (heatwaves) are the most relevant to forecast correctly. Having said that, it is obvious that only the move to a risk-based framing with probabilistic predictions allows to address the fundamental problem.

I am fully aware that this is a somewhat sweeping comment that can't be easily addressed without much further work. Therefore, I suggest the authors at least discuss the shortcomings of the current approach and how these could be overcome. One option that would be similar in complexity, but inherently probabilistic, would be to forecast threshold exceedance probabilities with a logistic regression model. Obviously, the evaluation would need to be adjusted accordingly.

We thank the reviewer for their insightful comment on a crucial aspect of seasonal forecasting, namely the need for probabilistic information. Only recently have probabilistic data-driven approaches emerged in the field of short-term forecasting (Price et al., 2025); such an advancement would be a huge step forward for data-driven seasonal forecasting as well, and should be a key objective for future versions of our proposed system. Aside from purely data-driven approaches, improved probabilistic information may be achieved by hybrid approaches, as proposed below. We mention these shortcomings in the Discussion.

Lines 237-240: “A crucial difference to the dynamical systems is that the current data-driven approach is inherently deterministic; future efforts must either explore probabilistic alternatives [54] or how to leverage the use of a single skillful members, such as for dynamical ensemble member selection [50].”

Regarding the loss function, we chose RMSE normalised (N-RMSE) by the interannual variability in order to alleviate the issue of smoothing and regression to the mean. N-RMSE is strongly related to correlation, itself a key skill metric for seasonal forecasting and validation of extreme indices. The authors are aware of other skill scores, or potential loss function, for extremes but these are usually applied to binary datasets (e.g. the Symmetric Extremal Dependence Index - SEDI - score).

Price, I., Sanchez-Gonzalez, A., Alet, F., Andersson, T.R., El-Kadi, A., Masters, D., Ewalds, T., Stott, J., Mohamed, S., Battaglia, P. and Lam, R., 2025. Probabilistic weather forecasting with machine learning. Nature, 637(8044), pp.84-90.

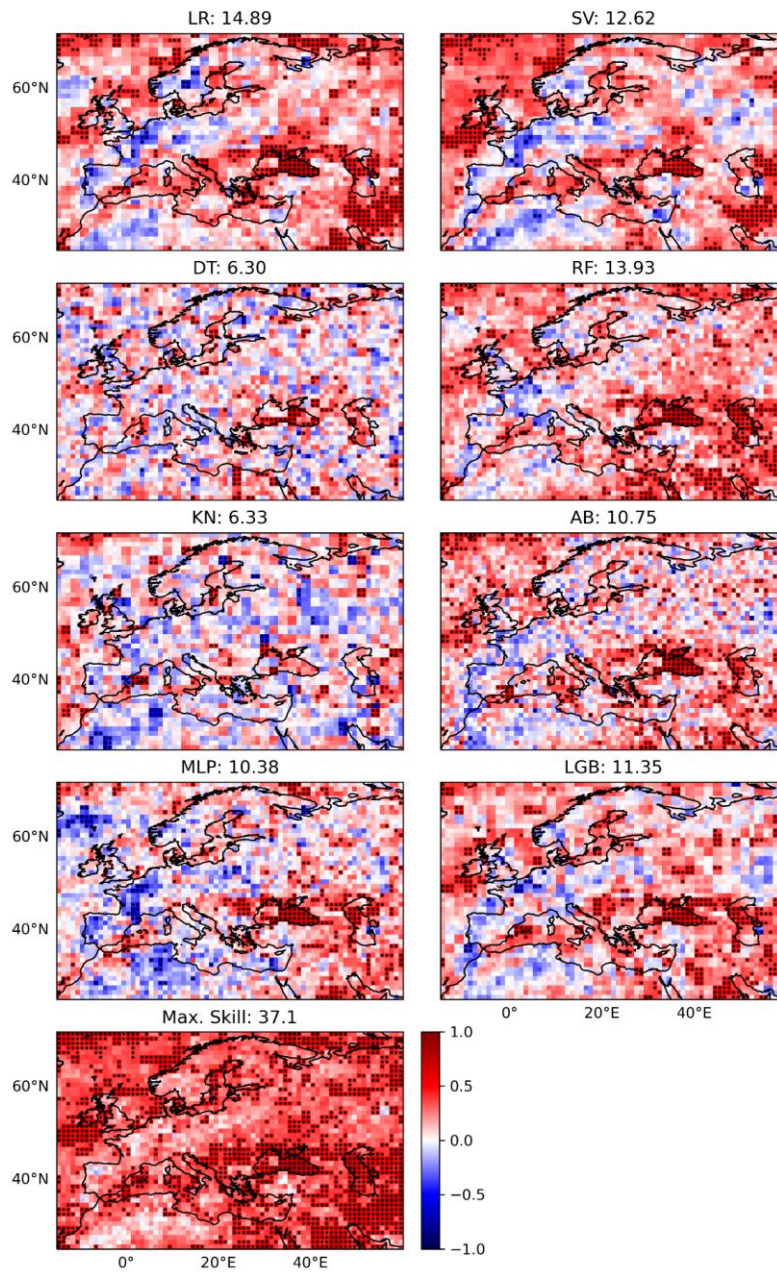
Model selection

The authors fail to motivate why using an evolutionary approach for feature selection is necessary. Arguably, the same result could be achieved with well-established model selection approaches for linear models that are computationally way cheaper (e.g. forward/backward stepwise selection). As there is a plethora of algorithms for model selection available, I encourage the authors to better motivate their choice.

We thank the reviewer for highlighting the need to better justify the choice of algorithm.

Forward stepwise selection is a brute-force method. For the optimisation algorithm applied here, one CPU hour is required for optimisation on a grid point with our approach; it is unlikely that a method that loops through all possible combinations of predictors to be faster than this. Here, we provide a comparison to a cheaper correlation-based feature selection. We identify the predictors which are significantly correlated (either positively or negatively) to the target dataset (NDQ90 in ERA5 from 1993-2016), for each grid point. Correlations are determined for each time-lag. These act as the selected features. We find that the skills of the resulting forecasts do not match those of the optimisation-based forecasts (Plot 2.2). We have added this to the discussion.

Lines 211-217: “To illustrate the benefits of the optimisation-based method, we compare to an alternative and simpler method of feature selection based on linear correlation analysis, in which the selected predictors are those significantly (positively or negatively) correlated to the target data during the period 1993-2016. Although the correlation approach selects a similar number of predictors (7.8% on average across the domain), the resulting forecasts are considerably poorer in skill (Ext. Data Fig. 11). This highlights the greater ability of our optimisation-based approach to capture physically plausible predictors compared to simpler statistical-based approaches.”



Plot 2.2. Skill score (correlation) of forecasts of May-July NDQ90 in the data-driven approach using a range ML models over the forecast test period 1993-2016, for the correlation-based feature selection. The following regression ML models are used: Linear Regression (LR), Support Vector (SV), Decision Tree (DT), Random Forest (RF), K-nearest neighbour (KN), AdaBoost (AB), Multi-Layer Perceptron (MLP), Light Gradient Boost (LGB). The skill corresponding to the best model in each grid point, used for the data-driven approach (Fig. 5) is shown by Max. Skill. Black stippling represents statistically significant correlation at the 95% confidence interval. Numbers in the titles correspond to the percentage of the domain covered by significant skill.

Minor comments

L89: the first implication is trivial and could be avoided. For the second implication, please elaborate how the similarity of the skill pattern with dynamical models demonstrates the limitations of the approach.

We have now rephrased this section.

Line 95-98: *“By applying the data-driven approach first to the model world, we isolate where the use of reduced-dimensionality predictors provides insufficient predictability. If the framework cannot recreate the paleo-climate model training data, then it is highly unlikely to perform well on the test data or in the real-world forecasting.”*

Figure 1: Why are there random under-performing solutions after 1000s of optimization steps far removed from the bulk of solutions that show a very predictably development (i.e. the orange red dots in the right part of the relative point clouds)?

While the optimisation algorithm converges on a particular solution, in these later stages of the optimisation the algorithm occasionally tests alternative combinations to ensure the convergence is indeed correct. Hence there are, as the reviewer suggests, random solutions which underperform. The algorithm however quickly corrects, hence the small number of these types of points. This behaviour is a means of testing the robustness of the solution converged to.

Figure 2 suggests that the method suffers from overfitting to the training data, as the performance on the test set is considerably worse compared with the training set. Is there any control of optimization to stop before such overfitting entails? If not, I suggest to split the training period in a training and validation period and use the latter to stop the optimization when overfitting occurs (i.e. when training and validation scores markedly start to differ).

We thank the reviewer for raising this point; we employ a 5-fold cross-validation of training period to reduce overfitting, but mistakenly did not previously state this in the manuscript. It has now been added to the Methods.

Lines 338-339: *“The training score is calculated with 5-fold cross-validation of the training period to reduce overfitting”.*

Figure 2: I would love to see the same metric also for the ERA5 based test set. Not least to get an impression of the differences between past2k and ERA5.

Given the poor training achieved in 50 years (Plot 2.1, Ext. Data Fig. 12) - the equivalent available in ERA5 - we would not expect the optimisation to accurately capture the predictors in ERA5. Another key issue is that the HW index is defined relative to the 1993-2016 baseline period, as per the dynamical systems. This means there are many years of no HW days in the early half of the record, and would hinder any training.

Figure 4/5: Field significance of results: clearly there is no indication of significant differences. Under the null hypothesis (and iid assumptions), you would expect at least 5% of the grid boxes to show significant differences. With correlation between grid boxes, you could even get a much larger fraction under the null.

We thank the reviewer for raising this point, and should discuss the results more carefully.

We have reworded the Results analysis.

Lines 151-154: *“When compared to the individual systems (Fig. 5), the data-driven approach is more skillful over certain areas, such as over Eastern Europe when compared to CMCC-35 and ECMWF-51, the previously mentioned Scandinavian “corridor” in MF-8, and over France in DWD-21. However, the skill increase is rarely statistically significant.”*

L224: typesetting error in units

We thank the reviewer for spotting this, and have corrected the error.

L246: It is not clear to me how NDQ90 for the dynamical systems and the past2k simulation has been computed. Are values normalized using quantiles estimated from ERA5 or from their own ‘climatology’?

We thank the reviewer for spotting this missing information. Each product uses its own respective climatology, thereby adjusting the mean biases inherent to the dynamical seasonal prediction systems.

Lines 289-297: *“The target data used in this study is the number of days, from 1st May to 31st July, in which TMX exceeds the climatological 90th percentile (NDQ90). Each product uses its own respective climatology, thereby adjusting the mean biases inherent to the dynamical seasonal prediction systems. For ERA5, the 90th percentile is calculated by first averaging each calendar day over the 1993-2016 period and then applying an 11-day running mean to smooth the climatology [62]. For past2k, the period used in the last 30 years of the simulation (1821-1850), a choice which is justified by the stability of the climate throughout the simulation [61]. The NDQ90 index provides a measure of the propensity of a season to experience extreme temperatures, and displays similar variability and predictability to other indices based on intensity [8, 18]”.*

L294: I assume the different data-driven approaches that are ultimately applied to ERA5 are also fitted on the training set of past2k (0-1600). Could you please clarify and state the setup more clearly?

For the data-driven forecasts, we train on the full 1850 year period. The feature-selection uses a training period of 1600 years. We have added this to the Methods.

Lines 354-357: *“The final step is to apply the method to real-world data. This requires simply changing the test period to 1993-2016 and the test predictors to those of ERA5, and extending the training period to cover the full 1850 year period of past2k. Given that past2k and ERA5 have different grids, a nearest-neighbour mapping function is used to associate past2k grid cells with those of ERA5.”*

L328: In the interest of reproducibility, I encourage the authors to make the pre-processed features and targets available as an analysis-ready dataset.

We agree fully with this suggestion. We have created a github repository which contains:
1 - files used to run the framework on a European domain
2 - Scripts to download the target and predictor data from past2k and ERA5.
3 - More intuitive python notebooks which go through the framework step-by-step, ideal for use in training and teaching.

In the case of acceptance for publication, we would prepare a zenodo folder with the final version of the target and predictor data.

We have added the links to the Code Availability statement.

Lines 399-406: *“The optimisation based feature selection framework and seasonal forecasting framework is freely available at <https://github.com/climateintelligence/DDHWSF.git>. The repository contains the scripts used to perform the optimisation and forecasting in this paper, and also step-by-step Jupyter notebooks used for training.”*

Reviewer #3 (Remarks to the Author):

In this manuscript, the authors used the optimization-based feature selection method to select different climate variables and constructed a prediction model based on regression and machine learning methods, focusing on the seasonal prediction of European heatwaves. The manuscript has reasonable logic and clear expression; however, I have the following major concerns that require revision.

Abstract needs to be improved, for example, skill improvement and what the key indicators are should be clearly stated.

The abstract has been updated.

Line 261-262: The clustering method with $k=5$ may not be able to completely separate the modes of different regions and different variables. Using Varimax-EOF for modal decomposition may be more helpful for feature extraction. If possible, the author can consider comparing the performance with other dimensionality reduction methods.

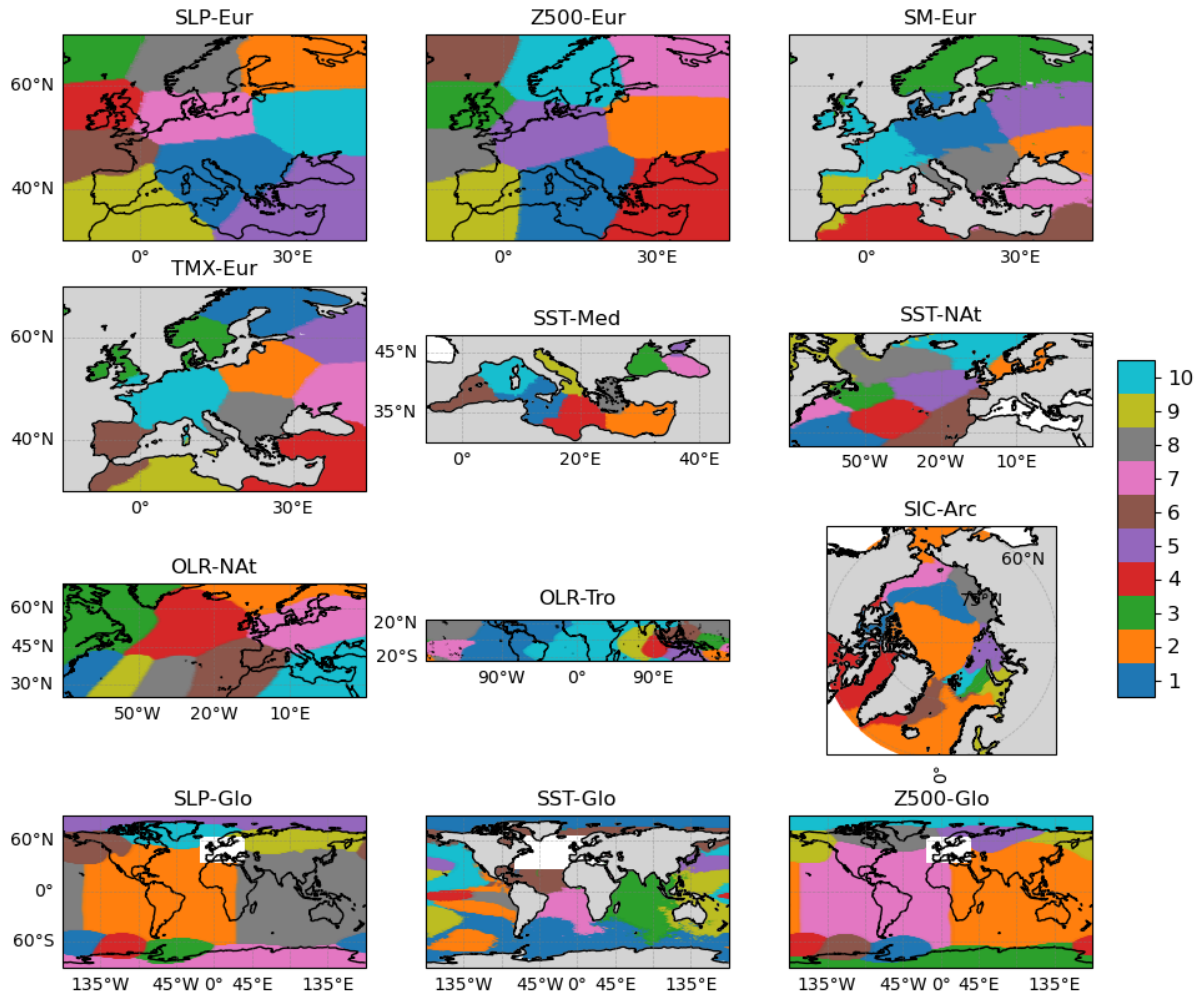
We thank the reviewer for raising this point. We do not use EOFs for several reasons. First, they reduce accessibility to spatial information. Second, there is the issue of “blending” of modes. Lastly, there is the arbitrary choice of EOFs to include. Arguably, given the rarity of extreme events, they are linked to the EOFs which explain much less of the total variability. As such, it would be necessary to include dozens of EOFs for each variable or domain, thereby increasing the computational expense of the feature selection.

$k=5$ is an arbitrary choice made to demonstrate the capability of the framework with large dimensionality reduction. The framework permits the use of any potential predictors the user wishes to study. We add the following line to the Methods:

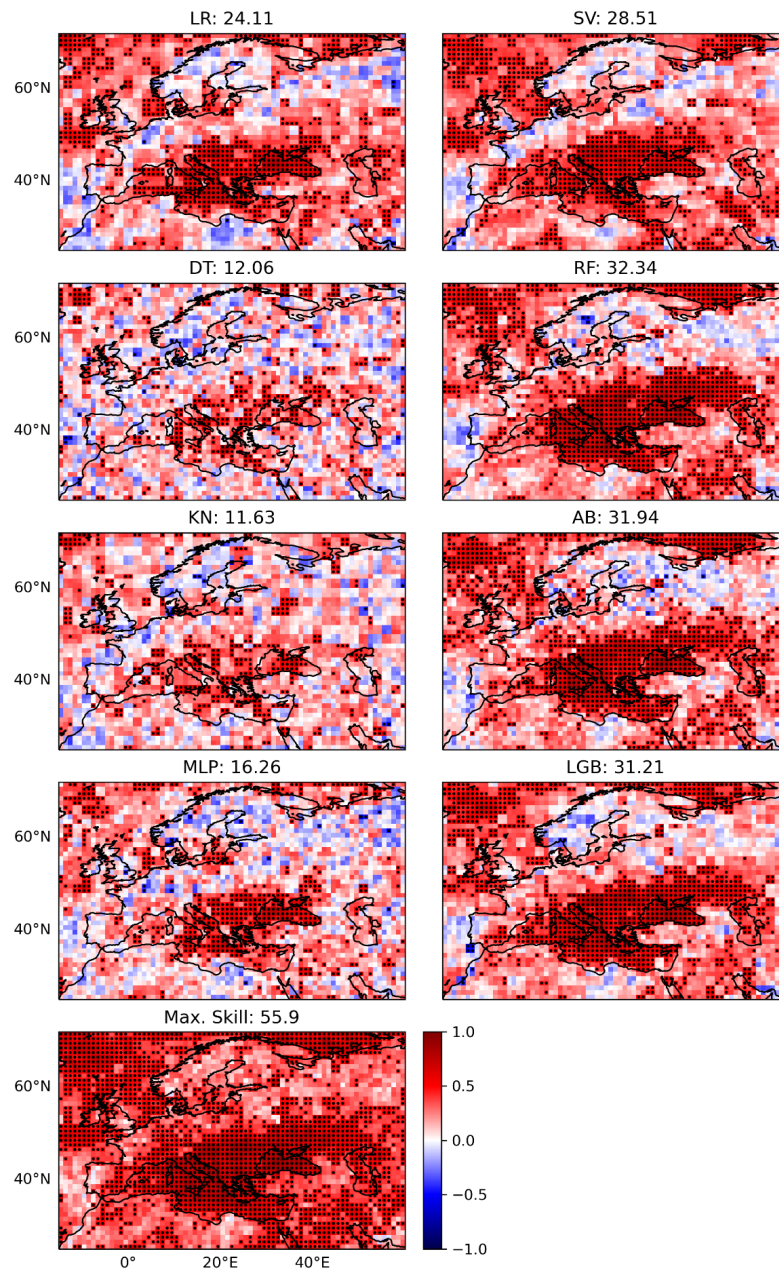
Lines 320-323: *“Given that our objective is to demonstrate the capability of such a framework, we use $k=5$ as a compromise between choosing a large pool of predictors and performing a suitable reduction of the dimensionality of the problem.”*

To test the sensitivity of results to the choice of potential predictors, we have now also performed an equivalent experiment using $k=10$ for each variable-domain (i.e. doubling the number of predictors; Plot 3.1). The optimisation took roughly 2-3 as many CPU hours as for the case in which $k=5$. The resulting data-driven seasonal forecasts are shown in Plot 3.2, and display negligible differences in skill to the experiment with $k=5$ clusters (Ext. Data Fig. 8). Thus, there is apparently little sensitivity to the cluster number.

The lack of added skill when using more potential predictors implies that the large-scale information gained from training/optimising on the paleo simulation is sufficiently represented by $k=5$, and there is no clear added-value provided by choosing $k=10$. Performing an extensive analysis on dimensionality reduction is not the scope of this paper, and thus we choose not to include this analysis in the paper.



Plot 3.1. Clusters of variables used to construct time series of potential predictors. K-means clustering is applied to ERA5 time series of anomalies for the period 1981-2010, with $k=10$.



Plot 3.2. Skill score (correlation) of forecasts when using increased number of clusters ($k=10$) as predictors.

Line 276-277: The mode integration results used by the author start from the 0th year. Did the author consider the spin-up problem? Are the mode integration results from the 0th to the 100th year stable? This will directly affect the quality of the model simulation.

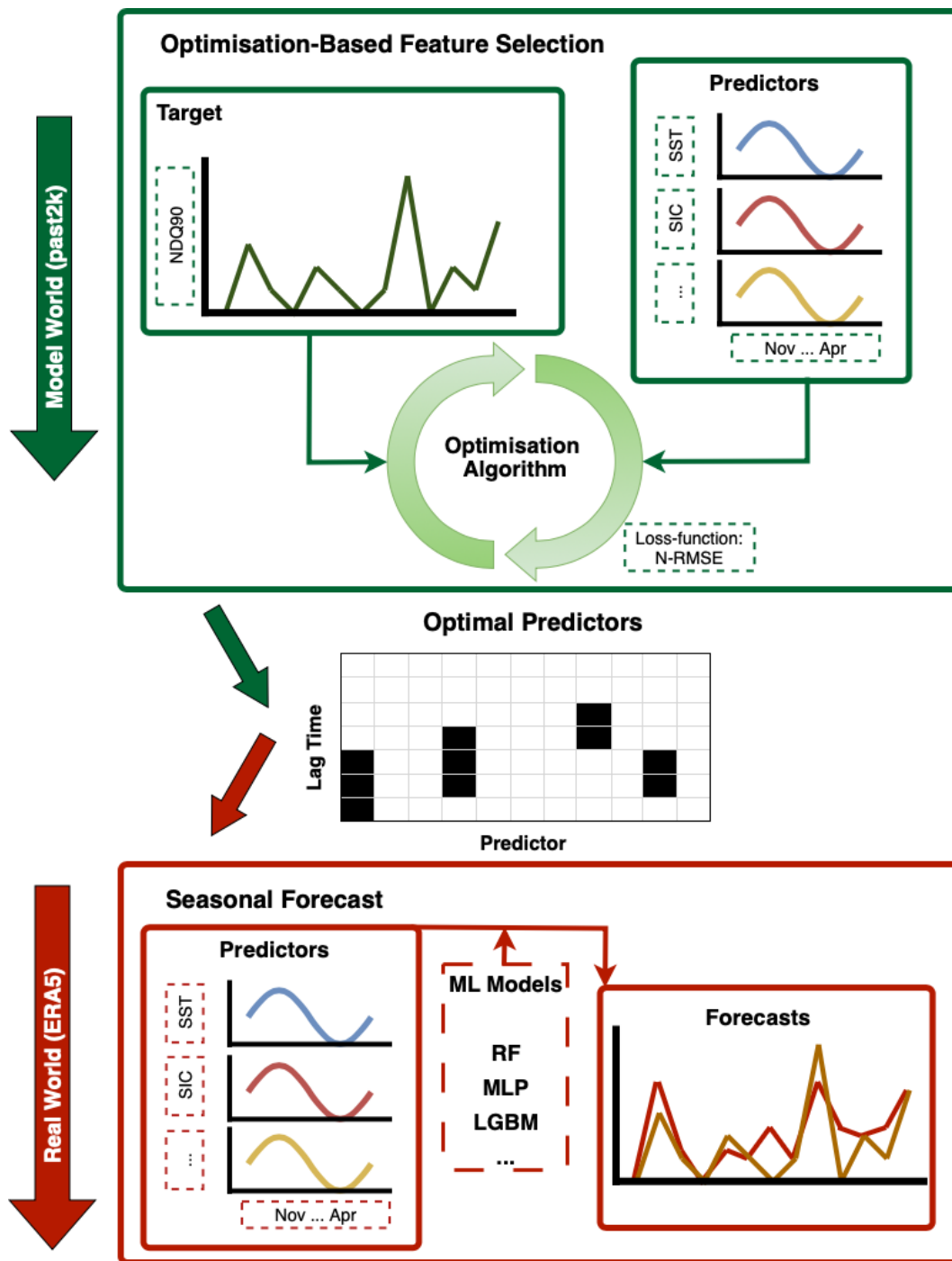
We thank the reviewer for pointing out this missing information. The spin-up time was 1200 years, prior to the year 0 and the time period used for training. We have added the following reference and text:

van Dijk, E., Jungclaus, J., Lorenz, S., Timmreck, C. and Krüger, K., 2022. Was there a volcanic-induced long-lasting cooling over the Northern Hemisphere in the mid-6th–7th century?. *Climate of the Past*, 18(7), pp.1601-1623.

Lines 258-259: *“The spin-up time of the simulation was 1200 model years prior to the year 0.”*

Line 247-317: The author can add a schematic diagram of the model structure and experimental process to help readers understand MMEA.

We agree with this suggestion and have prepared a schematic explaining the optimisation-based feature selection and the seasonal forecasting, and have added it to the Extended Data Figure 1 (Plot 3.3).



Plot 3.3. Schematic of the optimisation-based feature selection framework for seasonal forecasts of heatwave indices. The feature selection is performed in the model world (green), and the corresponding features in ERA5 are used as input into ML models to make the forecasts. Variables in dashed boxes represent adaptable components of the framework: the target data, predictors and time-lags, loss function, corresponding predictors in the real-world data, and the choice of ML models used to make the forecasts.

Line 290-291: The author emphasized in the title that the main work of this article is feature selection, but the details of the feature selection part were not elaborated, such as what the

optimization is. Is the loss function reduced by backpropagation, or is the hyperparameter adjusted randomly? The author should briefly introduce the optimization method in the article.

We thank the reviewer for highlighting this important point and have improved the introduction and description of the methods.

Lines 324-332: *“In the second step, the identification of predictors of extreme summers is treated as an optimisation problem. The optimisation algorithm employed is the Probabilistic Coral Reef Optimization algorithm with Substrate Layers (PCRO-SL) [44], which uses a multi-model method to combine different search procedures. In particular, it has recently been adapted to create a Spatio-Temporal Cluster-Optimized Feature Selection (STCO-FS) for heatwaves [63]. While previously the STCO-FS has been applied to detection of HWs, here it is adapted to the seasonal forecasting. We provide a description of the seasonal forecast set-up of STCO-FS and refer readers to aforementioned references for more complete technical descriptions of the optimisation algorithm PCRO-SL and feature selection set-up STCO-FS”.*

The hyperparameters of the optimisation algorithm are set by trial and error, based on experiences in previous studies finding the compromise between computation efficiency and quality of output (Perez-Aracil et al., 2023), showing low variability between different configurations. Performing a sensitivity analysis of hyperparameters is an enormous task beyond the scope of this study; besides, the current set-up is clearly capable of providing optimised predictors for the problem.

As the model is not based on a neural network, we have not implemented backpropagation in the framework.

The work which shows the methodology has been accepted for its publication in the journal *Weather and Climate Extremes*. The pre-print version of this work can be found here: <https://arxiv.org/abs/2411.10108>

Ext Data Fig.2 ‘Cluster maps are shown in Fig. X.’: Missing the specific image number.

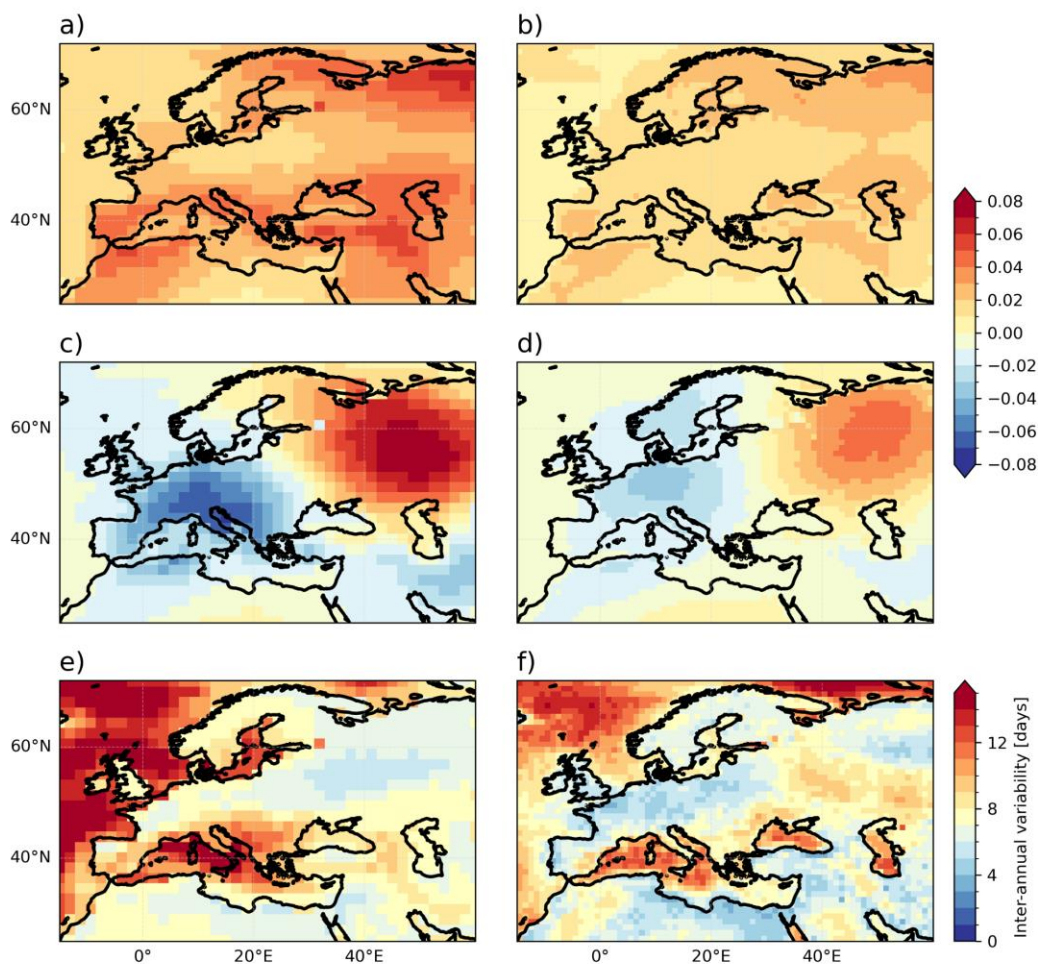
We thank the reviewer for spotting this error, and have added the correct number.

Line 213: The author only selected the simulation results of one model, but did not evaluate the simulation of MPI-ESM paleo-simulation. It is recommended that the author supplement the evaluation results of the model to enhance the credibility of subsequent conclusions.

We thank the reviewer for this suggestion. We now include maps of inter-annual variability of NDQ90 (ERA5 1993-2016 and past2k 1601-1850) and the EOF corresponding to the first principal component for both (Ext. Data Figure 5, Plot 3.4). Variability explained by the first EOF is 54 and 48 % for past2k and ERA5 respectively, with the

second EOFs contributing to 6% and 9% respectively. A comparison has been added to the Methods section.

Lines 268-274: *“The representation of summer temperature and heatwaves in past2k is found to agree with ERA5 (Ext. Data Fig. 5). First, the patterns of interannual variability of NDQ90 are similar across Europe, although the magnitude is consistently higher in past2k. For example, peaks appear across the Mediterranean basin, the Caucasus and north-western Russia. In the two leading Empirical Orthogonal Functions (EOFs) of average summer TMX, past2k and ERA5 resemble each other in terms of patterns and magnitude; the leading EOF is dominated by variability over Russia while the second clearly separates land and ocean variability.”*



Plot 3.4. Comparison of temperature and heatwave statistics in ERA5 and past2k. Interannual variability (calculated as the standard deviation) of the number of May-July days exceeding 90th percentile (NDQ90, target data for study) in (a) past2k over the test period 1601-1850 and (b) ERA5 over 1993-2016. Variability explained by the first EOF is 54 and 48% for past2k and ERA5 respectively, with the second EOFs contributing to 6% and 9% respectively.

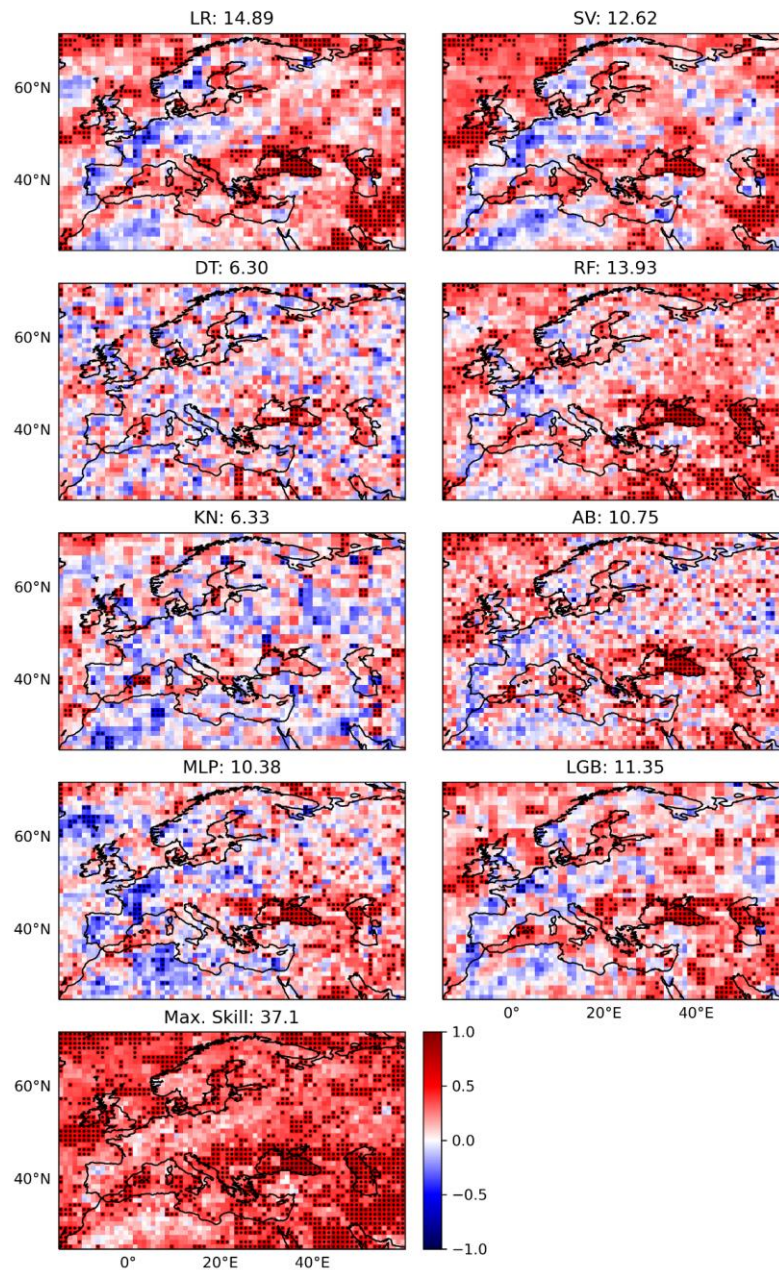
It is recommended that the author discuss the comparison between the feature selection method in this article and other commonly used feature selection methods (Pearson correlation coefficient method, distance correlation coefficient method, etc.) to reflect the advantages of this method.

Our response to this point can be found below.

As far as I understand, the major content of the manuscript focuses on seasonal predictions of European heat waves. Although one feature selection method is used, the author lacks comparison with other feature selection methods. Therefore, I think the title of this article can be modified accordingly unless the authors put more efforts on the feature selection method.

We agree this can aid in justifying the use of our proposed method, and respond to the previous two points together. As a simpler alternative to the evolutionary algorithm, we perform a correlation-based feature selection. We identify the predictors which are significantly correlated (either positively or negatively) to the target dataset (NDQ90 in ERA5 from 1993-2016), for each grid point. Correlations are determined for each time-lag. These act as the selected features. We find that the skill of the resulting forecasts does not match those of the optimisation-based forecasts (Ext. Data Fig. 11, Plot 3.5). This analysis has been added to the Discussion.

**Lines 206-217: “A benefit of the optimisation algorithm used is206
its capacity to filter out unnecessary predictor information and identify the key
predictors, thereby allowing the input of many potential predictors. Across the domain
studied here, between 3.2 and 10.6% (6.5% on average) of the cluster-lag combinations
are selected as predictors. Generally, including more predictors slows down ML
forecasts, ruling out the possibility of including all 2000 in this case. To illustrate the
benefits of the optimisation-based method, we compare it to an alternative and simpler
method of feature selection based on linear correlation analysis, in which the selected
predictors are those significantly (positively or negatively) correlated to the target data
during the period 1993-2016. Although the correlation approach selects a similar number
of predictors (7.8% on average across the domain), the resulting forecasts are
considerably poorer in skill (Ext. Data Fig. 11). This highlights the greater ability of our
optimisation-based approach to capture physically plausible predictors compared to
simpler statistical-based approaches.”**



Plot 3.5. Skill score (correlation) of forecasts of May-July NDQ90 in the data-driven approach using a range ML models over the forecast test period 1993-2016, for the correlation-based feature selection. The following regression ML models are used: Linear Regression (LR), Support Vector (SV), Decision Tree (DT), Random Forest (RF), K-nearest neighbour (KN), AdaBoost (AB), Multi-Layer Perceptron (MLP), Light Gradient Boost (LGB). The skill corresponding to the best model in each grid point, used for the data-driven approach (Fig. 5) is shown by Max. Skill. Black stippling represents statistically significant correlation at the 95% confidence interval. Numbers in the titles correspond to the percentage of the domain covered by significant skill.

The prediction skills of the data-driven method in this paper are comparable to those of the dynamical prediction models. Does this mean that the feature selection method is not advantageous? Compared with the high prediction skills obtained by previous end-to-end data-driven methods (Chattopadhyay et al., 2020, etc.), the author may want to add discussions on this.

Reference:

Chattopadhyay, A., Nabizadeh, E., & Hassanzadeh, P. (2020). Analog forecasting of extreme-causing weather patterns using deep learning. *Journal of Advances in Modeling Earth Systems*, 12(2), e2019MS001958.

We thank the reviewer for raising this point. There are areas in which our data-driven system outperforms dynamical models (Fig. 5), although rarely significantly. The current state-of-the-art of data-driven models in the domain of weather forecasting outperforms dynamical systems in some instances but not in others (see, for example, the study on extremes by Olivetti & Messori, 2024).

Olivetti, L. and Messori, G., 2024. Do data-driven models beat numerical models in forecasting weather extremes? A comparison of IFS HRES, Pangu-Weather, and GraphCast. Geoscientific Model Development, 17(21), pp.7915-7962.

The FS method is advantageous in places because DD skill is slightly higher than dynamical skill, for example. In short-term forecasting, there are examples of DD approaches outperforming dynamical approaches. Here, in the seasonal forecasting example, there is also evidence of DD outperforming. However, in seasonal forecasting, given the relatively small number of samples (24 years), caution must be exercised when interpreting increases in skill, by referring to statistical significance as we do here.

A crucial added-value, besides computational resources or differences in skill, is the scientifically-relevant information derived from the feature selection.