nature portfolio

Peer Review File

Detecting marine heatwaves below the sea surface globally using dynamics-guided statistical learning

Corresponding Author: Professor Zhao Jing

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 Professor Jing,

Your manuscript titled "Detecting Marine Heatwaves below the Sea Surface Globally Using Physics-Informed Statistical Learning" has now been seen by 3 reviewers, and we include their comments at the end of this message. They find your work of interest, but some important points are raised. We are interested in the possibility of publishing your study in Communications Earth & Environment, but would like to consider your responses to these concerns and assess a revised manuscript before we make a final decision on publication.

We therefore invite you to revise and resubmit your manuscript, along with a point-by-point response that takes into account the points raised. Please highlight all changes in the manuscript text file.

Please submit your point-by-point 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.

Important: The response to reviewers must not include any figures, tables or graphs. If you wish to respond to the reviewer reports with additional data in one of these formats, please add them to the main article or Supplementary Information, and refer to them in the rebuttal. Due to current technical limitations, any figures, tables, or graphs embedded in your rebuttal will not be included in the peer review file, if published.

We are committed to providing a fair and constructive peer-review process. Please don't 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 referees' comments (which should be in a separate document to any cover letter), a tracked-changes version of the manuscript (as a PDF file) and the 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 **

We hope to receive your revised paper within six weeks; please let us know if you aren't able to submit it within this time so that we can discuss how best to proceed. If we don't hear from you, and the revision process takes significantly longer, we may close your file. In this event, we will still 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.

Please do not hesitate to contact us if you have any questions or would like to discuss these revisions further. We look forward to seeing the revised manuscript and thank you for the opportunity to review your work.

Best regards,

Dr Alireza Bahadori
Associate Editor

Communications Earth & Environment

EDITORIAL POLICIES AND FORMATTING

We ask that you ensure your manuscript complies with our editorial policies. Please ensure that the following formatting requirements are met, and any checklist relevant to your research is completed and uploaded 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

Furthermore, please align your manuscript with our format requirements, which are summarized on the following checklist:
Communications Earth & Environment formatting checklist

and also in our style and formatting guide Communications Earth & Environment formatting guide .

*** DATA: Communications Earth & Environment endorses the principles of the Enabling FAIR data project (http://www.copdess.org/enabling-fair-data-project/). We ask authors to make the data that support their conclusions available in permanent, publically accessible data repositories. (Please contact the editor if you are unable to make your data available).

All Communications Earth & Environment manuscripts must include a section titled "Data Availability" at the end of the Methods section or main text (if no Methods). More information on this policy, is available at http://www.nature.com/authors/policies/data/data-availability-statements-data-citations.pdf.

In particular, the Data availability statement should include:
- Unique identifiers (such as DOIs and hyperlinks for datasets in public repositories)
- Accession codes where appropriate
- If applicable, a statement regarding data available with restrictions
- If a dataset has a Digital Object Identifier (DOI) as its unique identifier, we strongly encourage including this in the Reference list and citing the dataset in the Data Availability Statement.

DATA SOURCES: All new data associated with the paper should be placed in a persistent repository where they can be freely and enduringly accessed. We recommend submitting the data to discipline-specific, community-recognized repositories, where possible and a list of recommended repositories is provided at http://www.nature.com/sdata/policies/repositories.

If a community resource is unavailable, data can be submitted to generalist repositories such as figshare or Dryad Digital Repository. Please provide a unique identifier for the data (for example a DOI or a permanent URL) in the data availability statement, if possible. If the repository does not provide identifiers, we encourage authors to supply the search terms that will return the data. For data that have been obtained from publically available sources, please provide a URL and the specific data product name in the data availability statement. Data with a DOI should be further cited in the methods reference section.

Please refer to our data policies at http://www.nature.com/authors/policies/availability.html.

REVIEWER COMMENTS:

Reviewer #1 (Remarks to the Author):

This work is an interesting attempt to combine physical oceanography and statistics to detect subsurface temperature anomalies, based on which subsurface Marine Heatwaves (MHWs) can be further identified. The physical oceanographic framework is the Quasi-Geostrophic (QG) dynamics used in earlier works to infer ocean’s interior based on surface data. By combining the QG dynamics and optimal prediction theory, the authors demonstrate that their constructed statistical model outperforms a traditional least-squares regression method and a machine-learning based approach. The manuscript is well written, and the structure is easy to follow. While there is a rapidly growing literature on MHWs, many studies have focused on surface MHWs, and understandings of subsurface MHWs have been lacking. In that regard, this is an important and timely work. I only have some minor comments for the authors to improve the manuscript. I envision recommending the publication of this work after these minor comments are addressed.

1. Because the nature of this work is largely methodology demonstration, the method section is key to this short article. While both the QG and statistical methods are not new, it is still necessary to describe the methodology more clearly. For example, description of steps in deriving subsurface temperature anomaly from sea surface temperature (SST) and sea surface height anomaly (SSHA) along with considerations and assumptions at each step would make the manuscript more accessible/acceptable to a wider audience.
2. Further discussion of the applicability of the method would be helpful to the readers. For example, QG is good for mesoscale and larger scale processes, but is not good for submesoscale inference. This may not be a problem for the current study, which implicitly targets mesoscale and larger MHWs due to the relatively coarse-resolution products used here, but it would make the paper well thought out if the authors expanded the discussion on this aspect, for example, within the context of Surface Water and Ocean Topography (SWOT) data.
3. Along the same line, more discussion of the varying performance in different geographic regions would make the text more appealing. In addition to salinity's role in the eastern Atlantic and southeastern Pacific, many marginal seas and the higher latitude oceans are subject to strong buoyancy/freshwater influence. Also, QG formulation is less ideal for coastal oceans, yet the MCC values are not much lower in the marginal seas. Does this imply that the statistical model is less sensitive to the physical information? Or this is due to the imperfect representation of coastal processes in these global products?

More minor comments:

Page 2, line 27: "They are occurring ubiquitously across the global ocean." Ubiquitously may be too strong. Suggest something like "occurring frequently across the global ocean in recent years".

Page 2, line 29: "in-situ" to "in situ" throughout.

Page 3, paragraph 2: Great discussion bringing up the importance of subsurface MHWs. Another example is a subsurface MHW in early 2017 on the Northeast US continental shelf, which had significant impact on the ecosystem and commercial fishing industry. The onset of this event involves both mesoscale and submesoscale processes.

Gawarkiewicz, G., K. Chen, J. Forsyth, F. Bahr, A. M. Mercer, A. Ellertson, P. Fratantoni, H. Seim, S. Haines, and L. Han, 2019: Characteristics of an advective marine heatwave in the Middle Atlantic Bight in early 2017. Frontiers in Marine Science, 6, 712.

Chen, K., G. Gawarkiewicz, and J. Yang, 2022: Mesoscale and Submesoscale Shelf-Ocean Exchanges Initialize an Advective Marine Heatwave. Journal of Geophysical Research: Oceans, 127, e2021JC017927.

Page 8, Figure 1: "Area where MCC is not significantly positive at the 95% confidence level is filled in white.". This confuses with regions that do not have actual data, e.g., 100m/200m depth level in the continental shelf regions. Suggest using a different color or symbol for low MCC values.

Page 12, line 244: "… since the true model is linear." It might confuse some readers into thinking this is saying the true ocean is linear.

Reviewer #2 (Remarks to the Author):

This study detects the marine heatwaves (MHWs) using the SST and SSHA. Predictors, SST and SSHA, are picked based on the QG equations, and to consider its spatiotemporal variations, authors formulated multiple linear regression model with geographically and seasonally varying coefficients (so-called QSVC). The performance of the QSVC model is compared to that of OLS model with temporally-fixed coefficient, and CNN model, and authors demonstrated that the estimation quality of the QSVC is systematically better than the other two models.
I am serious doubtful about recommending this article for publication in Communications earth & environment due to the following two reasons. First, there is a huge misleading in using the terminology 'physics-informed'. Normally, 'physics-informed' machine learning refers the machine learning with a custom-designed loss functon(Kashinath et al., 2021). That is, by adding the constraints based on the physical equation to the conventional loss function, the model's solution is regularized to keep the physical constraint. On the other hand, in this study, the physical constraint is not explicitly given to formulate their model, and the QG equations are used only for the predictor variable (i.e., SST and SSHA) selection. The usage of the physical equation in that way cannot refer to as a 'physics-informed' method. Second, it is highly expected that the multiple linear regression model with coefficients varying in space and time (i.e., QSVC model) would performs better

than that with coefficients varying only in space (OLS model). I do not see any new findings or insights about this point. Third, what is a purpose of estimating T' using SST? As there is no time-lag between the predictor and predictand, once MHWs at the surface is estimated using the current methods, the correlation should be 1 as far as I understand. That is, the implication of this study is not clear to me. Therefore, my recommendation is the reject.

References
Kashinath, K., Mustafa, M., Albert, A., Wu, J. L., Jiang, C., Esmaeilzadeh, S., ... & Prabhat, N. (2021). Physics-informed machine learning: case studies for weather and climate modelling. Philosophical Transactions of the Royal Society A, 379(2194), 20200093.

Reviewer #3 (Remarks to the Author):

The study by Zhang et al focuses on deriving the linear dynamical relationship between sea surface temperature (SST), sea surface height (SSH), and subsurface temperature, and building a seasonally-varying regression model with SST and SSH as predictors and subsurface temperature as predictand. By harnessing the linear relationship that varies by seasons, the authors show that this approach outperforms the simpler linear regression without considering seasonality, and also outperforms a convolutional neural network that learns a nonlinear relationship when the dynamics should be linear. I think it is a solid study, and I would recommend publication after revision.

My main comment is that the main texts lack key descriptive details, without which the readers have to go through all of the methods to grasp the key messages. The following is my main comment in detail:

Line 89-103: Thus far, the authors have introduced the dynamical and statistical learning methods used to infer subsurface temperature from the surface ocean variables. In this last paragraph of the introduction, the authors introduced their approach and remarked on its outperformance. However, I think that describing the authors' approach as "combining the ocean dynamics and optimal prediction theory" is far from sufficient. More descriptive details are necessary for readers to understand the fundamental basis of the model.

Line 117: Again description of the "optimal prediction theory" in the introduction would greatly help navigate the readers. This is especially useful for nature style papers. The authors should not expect readers to go to Method section for understanding the basic concepts of their methodology.

Line 219-223: The biggest difference between OLS and GSVC is that GSVC uses information from nearby spatial points within a short time range to derive the regression coefficients. These were smoothing techniques to alleviate the issue of small sample size, which is a significant issue since the authors want to derive coefficients for each calendar day in a year. Assuming there are enough samples, one could also derive the coefficients for each calendar day using OLS. I could also derive the coefficients for every 60 days and apply those coefficients for the center day (i.e., the number 30th day). So, I don't see that OLS and GSVC are fundamentally different. And I agree with the authors that the "OLS model is actually a reduced case of the GSVC model" or "GSVC model is an extension from OLS". I think it would be helpful to describe a bit more about the features of GSVC model in the main texts, especially highlighting the fact that it is a daily-resolution seasonally-varying model and that the OLS is not seasonally-varying. Similarly, if I understand correctly, the CNN model is also not seasonally-varying, i.e., the authors did not train the CNN separately for each season. This might be part of the reasons why CNN works poorly compared with OLS or GSVC.

Minor comments:

Line 46: Authors' names are missing in the reference #11.

Line 51: The following reference is also relevant:

Amaya, D. J., et al. (2023). "Bottom marine heatwaves along the continental shelves of North America." Nature Communications 14(1): 1038.

Line 109-112: Briefly elaborating the so-called "tight" relationship between SST, SSH and T' would be beneficial.

Line 153-159: Do you apply the same percentile threshold identified from the observation when determining if the subsurface temperature inferred from SST and SSH is a MHW?

Line 167-168: Figure 1 shows skill from -1 to 1 with color changing from blue to white to red. Areas where MCC is not significant are also filled in white. Thus, "white" color has different meanings in this figure. I would suggest to highlight areas where MCC is not significantly positive using hatching.

Line 183-184: Can the authors explain why the hit rate and false alarm rate are less effective metrics for assessing MHWs?

Line 185: "Further" should be replaced with "Furthermore".

Line 203-204: Can the authors clarify the meaning of "standard error"? Is it the range of globally-averaged MCC across 6 datasets?

Line 279: I am not quite sure if “physics-informed statistical learning method” would be the most accurate way to describe the proposed approach. I think it is a seasonally-varying linear regression model that built on the linear relationship of key ocean dynamics.

Line 305: I believe that the following should be the right reference for GLORYS.

Lellouche, J., et al. (2021). The Copernicus Global 1/12 Oceanic and Sea Ice GLORYS12 Reanalysis, Frontiers in Earth Science, 9.

Line 325-326: Why is “T_c” irrelevant to MHWs? Is it because “T_c” was removed from T to get T’? This sentence seems confusing to me.

Line 416: What is the use of the random noise in Eqn (13)?

Line 484: Can the authors explain why using a smaller weight on the “one” class and a larger weight on the “zero” class makes the model pay more attention to occurrence of MHW events? I thought a larger weight would mean a larger penalty for missing a prediction of a certain class. In the authors’ case, that would mean the model really don’t want to predict a wrong “zero”.

Line 574: I think it is more compelling to plot the figure within the range of 0 to 1.

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 Professor Jing,

Your manuscript titled "Detecting Marine Heatwaves below the Sea Surface Globally Using Dynamics-Guided Statistical Learning" 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 revise your paper one last time to address the remaining concerns of our reviewers. At the same time we ask that you 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 Research, 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,

Dr Alireza Bahadori
Associate Editor
Communications Earth & Environment

REVIEWERS' COMMENTS:

Reviewer #1 (Remarks to the Author):

The authors have addressed my previous comments. I find the revised manuscript acceptable for publication.

Reviewer #2 (Remarks to the Author):

To support authors' claim that the main contribution of their study that the outperformance of the GSVC over the OLS is meaningful, it should be demonstrated in the Introduction that the common approach in their research field is still based on the simple OLS that universally used for all seasons. I believe that this paper is worth publishing, only demonstrating that the OLS is still widely accepted in their research field.
(Personally, I don't believe their arguments that their findings, that the GSVC is better than the OLS, is valuable in their research fields. It is like a common sense to me, as it is already well recognized that the relationships between physical variables varies seasonally due to their seasonal cycle.)

Reviewer #3 (Remarks to the Author):

The authors have largely addressed my comments. And I really appreciate the effort. I have only a few very minor comments left. The first and foremost important one is that I think the equations should be moved back to the method section. I don't think simply moving the equations from the method section to the result section is going to help readers understand the methodology. I think both the first reviewer and I were asking for physical description of the method.

Other minor comments are as followed:

Line 53: I have no expertise in biology. But it seems weird to categorize salmons as "large organisms". Perhaps just use "fish" is sufficient.

Line 215-223: Perhaps the authors should explicitly mention how they treat climate change. Before training the model, do the authors detrend first so the system represents a stationary system?

Line 252-253: If I am not color blind, from my view, the land is in dark grey color, whereas insignificant areas are in light grey color?

Line 296-298: I wouldn't say that OLS regression and CNN do not "exploit ocean dynamics". They are still trying to capture the relationship between surface and subsurface ocean. They get reasonable results due to the fact that they were able to capture some of that relationship.

Line 359-365: I think that a fundamental issue is the lack of the long record for use in the more complicated methods. Perhaps the authors could comment on that the paragraph.

Reply to the first reviewer

We are very grateful to you for your time in carefully reading our manuscript and providing helpful comments that make our manuscript better. We have carefully considered each of your comments (in blue) and revised the manuscript accordingly. Please find our response (in black) to your comments below.

Reviewer #1 (Comments to Author):

This work is an interesting attempt to combine physical oceanography and statistics to detect subsurface temperature anomalies, based on which subsurface Marine Heatwaves (MHWs) can be further identified. The physical oceanographic framework is the Quasi-Geostrophic (QG) dynamics used in earlier works to infer ocean's interior based on surface data. By combining the QG dynamics and optimal prediction theory, the authors demonstrate that their constructed statistical model outperforms a traditional least-squares regression method and a machine-learning based approach. The manuscript is well written, and the structure is easy to follow. While there is a rapidly growing literature on MHWs, many studies have focused on surface MHWs, and understandings of subsurface MHWs have been lacking. In that regard, this is an important and timely work. I only have some minor comments for the authors to improve the manuscript. I envision recommending the publication of this work after these minor comments are addressed.

Thanks for your positive comments.

1. Because the nature of this work is largely methodology demonstration, the method section is key to this short article. While both the QG and statistical methods are not new, it is still necessary to describe the methodology more clearly. For example, description of steps in deriving subsurface temperature anomaly from sea surface temperature (SST) and sea surface height anomaly (SSHA) along with considerations and assumptions at each step would make the manuscript more accessible/acceptable to a wider audience.

In our original manuscript, we briefly describe the key points of our methodology in the main text and present detailed description of our methodology in the Method Section. Following your and Reviewer 3's advice, we have described our methodology in details in the main texts and stated clearly the considerations and assumptions at each derivation step. Please see Line 109-223.

2. Further discussion of the applicability of the method would be helpful to the readers. For example, QG is good for mesoscale and larger scale processes, but is not good for submesoscale inference. This may not be a problem for the current study, which implicitly targets mesoscale and larger MHWs due to the relatively coarse-resolution products used here, but it would make the paper well thought out if the authors expanded the discussion on this aspect, for example, within the context of Surface Water and Ocean Topography (SWOT) data.

Thanks for your comment. As you pointed out, the QG dynamics is not suitable for submesoscale motions. The strong nonlinearity inherent in submesoscale motions may invalidate the linear dependence of subsurface temperature anomalies on SSTA and SSHA at submesoscales. We have cautioned the readers about the limitation of our study in the revised manuscript. Please see Line 381-389.

3. Along the same line, more discussion of the varying performance in different geographic regions would make the text more appealing. In addition to salinity's role in the eastern Atlantic and southeastern Pacific, many marginal seas and the higher latitude oceans are subject to strong buoyancy/freshwater influence. Also, QG formulation is less ideal for coastal oceans, yet the MCC values are not much lower in the marginal seas. Does this imply that the statistical model is less sensitive to the physical information? Or this is due to the imperfect representation of coastal processes in these global products?

Thanks for raising this important question. The GSVC models appears to perform equally well in the coastal and open ocean. We suspect that the superficially equal performance may be partially due to the insufficient resolution of global reanalysis products that lead to imperfect representation of coastal processes. We have cautioned readers about this issue in the revised manuscript. Please see Line 385-389.

As to the good performance of GSVC in some high-latitude regions, it may result from the deep mixed layer there. Although there is strong buoyancy/freshwater influence in high-latitude regions, the deep mixed layer makes the subsurface temperature anomaly (provided it is located within the deep mixed layer) highly correlated to SSTA. One instance is the subpolar region in the North Atlantic where the MCC is high and the mixed layer is deep.

More minor comments:

Page 2, line 27: "They are occurring ubiquitously across the global ocean." Ubiquitously may be too strong. Suggest something like "occurring frequently across the global ocean in recent years".

Thanks for your comments. We have revised it as "occurring more and more frequently across the global ocean in recent years", see Line 27.

Page 2, line 29: "in-situ" to "in situ" throughout.

Revised.

Page 3, paragraph 2: Great discussion bringing up the importance of subsurface MHWs. Another example is a subsurface MHW in early 2017 on the Northeast US continental shelf, which had significant impact on the ecosystem and commercial fishing industry. The onset of this event involves both mesoscale and submesoscale processes.

Gawarkiewicz, G., K. Chen, J. Forsyth, F. Bahr, A. M. Mercer, A. Ellertson, P. Fratantoni, H. Seim, S. Haines, and L. Han, 2019: Characteristics of an advective marine heatwave in the Middle Atlantic Bight in early 2017. Frontiers in Marine Science, 6, 712.

Chen, K., G. Gawarkiewicz, and J. Yang, 2022: Mesoscale and Submesoscale Shelf-Ocean Exchanges Initialize an Advective Marine Heatwave. Journal of Geophysical Research: Oceans, 127, e2021JC017927.

Thanks for your useful advice and references. We have added these examples to highlight the occurrence of subsurface MHWs and their catastrophic ecological impact, please see Line 56-58 and references 23 and 24.

Page 8, Figure 1: "Area where MCC is not significantly positive at the 95% confidence level is filled in white.". This confuses with regions that do not have actual data, e.g., 100m/200m depth level in the continental shelf regions. Suggest using a different color or symbol for low MCC values.

Thank you for your advice. We have filled the land region in grey to distinguish the region where MCC is not significantly different from zero.

Page 12, line 244: "… since the true model is linear." It might confuse some readers into thinking this is saying the true ocean is linear.

Sorry for the ambiguous expression. We have deleted this expression in the revised manuscript.

Reply to the second reviewer

We are very grateful to you for your time in carefully reading our manuscript and providing helpful comments that make our manuscript better. We have carefully considered your comments (in blue) and revised the manuscript accordingly. Please find our response (in black) to your comments below.

Reviewer #2 (Remarks to the Author):

This study detects the marine heatwaves (MHWs) using the SST and SSHA. Predictors, SST and SSHA, are picked based on the QG equations, and to consider its spatiotemporal variations, authors formulated multiple linear regression model with geographically and seasonally varying coefficients (so-called QSVC). The performance of the QSVC model is compared to that of OLS model with temporally-fixed coefficient, and CNN model, and authors demonstrated that the estimation quality of the QSVC is systematically better than the other two models.

I am serious doubtful about recommending this article for publication in Communications earth & environment due to the following two reasons. First, there is a huge misleading in using the terminology ‘physics-informed’. Normally, ‘physics-informed’ machine learning refers the machine learning with a custom-designed loss function (Kashinath et al., 2021). That is, by adding the constraints based on the physical equation to the conventional loss function, the model’s solution is regularized to keep the physical constraint. On the other hand, in this study, the physical constraint is not explicitly given to formulate their model, and the QG equations are used only for the predictor variable (i.e., SST and SSHA) selection. The usage of the physical equation in that way cannot refer to as a ‘physics-informed’ method.

We are grateful to you for pointing out this important issue. The physics can inform the statistical learning in many ways. First, it can be used to improve the learning performance by adding the constraints based on the physical equation to the conventional loss function. Second, it can provide insights into the true relationship between the predictors and response, which guides the proposal of appropriate statistical learning methods.

We fully agree with you that the expression “physics-informed” has been widely used to refer to the first case. To avoid misleading statement, we have replaced the terminology “physics-informed statistical learning” in our original manuscript as "dynamics-guided statistical learning" to highlight the fact that the proposal of our statistical learning method is guided by ocean dynamics.

Second, it is highly expected that the multiple linear regression model with coefficients varying in space and time (i.e., GSVC model) would performs better than that with coefficients varying only in space (OLS model). I do not see any new findings or insights about this point.

With the increase of computational resources and data volume, there is a tendency to use more and more complicated machine learning models like neural networks in oceanic studies and other scientific fields. However, more complicated models, albeit more flexible, do not necessarily outperform the simpler ones due to the bias-variance trade-off issue. Developing models with the optimal degree of complexity is thus crucial to enhance the learning performance.

The optimal model relies on the true relationship between response and predictors and is thus problem-specific. In this study, we first combine ocean dynamics and optimal prediction theory in statistics to demonstrate a geographically and seasonally varying linear relationship between the subsurface temperature anomaly and its predictors (SSTA and SSHA), and then propose the GSVC model to capture such relationships. We compare the GSVC model with a simpler model (OLS) and a more complicated one (CNN). On the one hand, the outperformance of the GSVC over the OLS is because the OLS is too simple to capture the varying relationship. On the other hand, the outperformance of the GSVC over the CNN is because the CNN is ***overly complicated***. In particular, the CNN has a highly complicated model architecture to handle the nonlinear relationship, whereas the true relationship is linear.

We think there are three major contributions of our study. First, our study proposes a useful model (GSVC) to retrieve subsurface temperature anomaly based on observable predictors, i.e., SSTA and SSHA, enhancing the capacity to monitor subsurface MHWs. Second, our study demonstrates how the ocean dynamics can guide the statistical learning and improve its learning performance. In specific, the ocean dynamics provides insight into the true relationship between subsurface temperature anomaly and its predictors, which motivates us to propose the GSVC to capture such relationship. Third, our study suggests that complicated models do not necessarily outperform the simple ones (e.g., the CNN performs even worse than the OLS for first predicting subsurface temperature anomaly and then detecting MHWs), providing caution about the blind pursuit of model complexity in practice.

We have revised the discussion section to highlight the contributions of our study. Please see Line 359-380.

Third, what is a purpose of estimating T' using SST? As there is no time-lag between the predictor and predictand, once MHWs at the surface is estimated using the current methods, the correlation should be 1 as far as I understand. That is, the implication of this study is not clear to me. Therefore, my recommendation is the reject.

The MHWs can be identified based on the temperature anomaly. Currently, satellites provide global-coverage high-resolution temperature measurements at the sea surface, so that the **surface MHWs** are well monitored and studied. In contrast, the subsurface temperature measurements are obtained from in situ platforms and spatio-temporally very sparse. These sparse measurements cannot be used to detect the **subsurface MHWs**. In this study, we aim to predict the high-resolution subsurface temperature anomaly T' from the high-resolution sea surface temperature anomaly (SSTA) and sea surface height anomaly (SSHA) measured by satellites. Once the high-resolution T' is available, we can monitor the subsurface MHWs. It should be noted that SSTA is not always highly correlated to T' especially in the deeper region. In fact, the SSHA becomes a more important predictor than SSTA in predicting T' at 200 m (Figure 2d).

Reply to the third reviewer

We are very grateful to you for your time in carefully reading our manuscript and providing helpful comments that make our manuscript better. We have carefully considered each of your comments (in blue) and revised the manuscript accordingly. Please find our response (in black) to your comments below.

Reviewer #3 (Comments to Author):

The study by Zhang et al focuses on deriving the linear dynamical relationship between sea surface temperature (SST), sea surface height (SSH), and subsurface temperature, and building a seasonally-varying regression model with SST and SSH as predictors and subsurface temperature as predictand. By harnessing the linear relationship that varies by seasons, the authors show that this approach outperforms the simpler linear regression without considering seasonality, and also outperforms a convolutional neural network that learns a nonlinear relationship when the dynamics should be linear. I think it is a solid study, and I would recommend publication after revision.

My main comment is that the main texts lack key descriptive details, without which the readers have to go through all of the methods to grasp the key messages. The following is my main comment in detail:

Thanks for your positive comments. In our original manuscript, we briefly describe the key points of our methodology in the main text and present detailed description of our methodology in the Method Section. Following your and Reviewer 1's advice, we have described our methodology in details in the main texts. Please see Line 109-223.

1. Line 89-103: Thus far, the authors have introduced the dynamical and statistical learning methods used to infer subsurface temperature from the surface ocean variables. In this last paragraph of the introduction, the authors introduced their approach and remarked on its outperformance. However, I think that describing the authors' approach as "combining the ocean dynamics and optimal prediction theory" is far from sufficient. More descriptive details are necessary for readers to understand the fundamental basis of the model.

Thank you for your advice. This part is aimed to provide a very brief summary of major advances of our study. To avoid make the Introduction Section cumbersome, we provide detailed explanations of our methodology in the Result Section of the main text in the revised manuscript. Please see Line 109-223.

2. Line 117: Again description of the "optimal prediction theory" in the introduction would greatly help navigate the readers. This is especially useful for nature style papers. The authors should not expect readers to go to Method section for understanding the basic concepts of their methodology.

Thanks for your advice. We have provided description of the optimal prediction theory in the main text in the revised manuscript. Please see Line 192-203.

3. Line 219-223: The biggest difference between OLS and GSVC is that GSVC uses information from nearby spatial points within a short time range to derive the regression coefficients. These were smoothing techniques to alleviate the issue of small sample size, which is a significant issue since the authors want to derive coefficients for each calendar day in a year. Assuming there are enough samples, one could also derive the coefficients for each calendar day using OLS. I could also derive the coefficients for every 60 days and apply those coefficients for the center day (i.e., the number 30th day). So, I don't see that OLS and GSVC are fundamentally different. And I agree with the authors that the "OLS model is actually a reduced case of the GSVC model" or "GSVC model is an extension from OLS". I think it would be helpful to describe a bit more about the features of GSVC model in the main texts, especially highlighting the fact that it is a daily-resolution seasonally-varying model and that the OLS is not seasonally-varying. Similarly, if I understand correctly, the CNN model is also not seasonally-varying, i.e., the authors did not train the CNN separately for each season. This might be part of the reasons why CNN works poorly compared with OLS or GSVC.

Following your advice, we have highlighted in the main text that the GSVC is a daily-resolution seasonally varying regression coefficient model and the OLS assumes a temporally constant regression coefficient. We have also explained to readers that the capacity of the GSVC for estimating seasonally varying regression coefficient is own to the fact that it borrows samples from nearby spatial points to enlarge sample sizes for estimation. Please see Line 211-214 and 298-301.

As you pointed out, we did not train the CNN separately for each calendar day. There are two reasons. First, if we trained the CNN separately for each calendar day, there would only be about 20 snapshots (samples) to train each model, which may result in severe overfitting. Second, training 365 CNN models causes unaffordable computational burden and is infeasible in practice. We agree with you that the poor performance of CNN compared to GSVC is partially due to the incapacity of CNN to capture the seasonally varying relationship. However, as both CNN and OLS adopt a temporally constant relationship, the poor performance of CNN compared to OLS should be primarily attributed to the fact that the true relationship is linear.

Minor comments:

Line 46: Authors' names are missing in the reference #11.

Thanks for pointing out this issue. Revised.

Line 51: The following reference is also relevant:

Amaya, D. J., et al. (2023). "Bottom marine heatwaves along the continental shelves of North America." Nature Communications 14(1): 1038.

Thanks for your advice and useful reference. We have added the example and reference in the revised manuscript. Please see Line 61-62 and Reference 26.

Line 109-112: Briefly elaborating the so-called "tight" relationship between SST, SSH and T' would be beneficial.

Thank you for your advice. We demonstrated this "tight" relationship in details in the later section. Please see Line 115-203.

Line 153-159: Do you apply the same percentile threshold identified from the observation when determining if the subsurface temperature inferred from SST and SSH is a MHW?

Yes, when detecting MHWs based on T' predicted from SSTA and SSHA, the associated threshold θ' is computed from the "true" T' in the reanalysis dataset. In the revised manuscript, we have added the above statement in the Method Section. Please see Line 426-428.

Line 167-168: Figure 1 shows skill from -1 to 1 with color changing from blue to white to red. Areas where MCC is not significant are also filled in white. Thus, "white" color has different meanings in this figure. I would suggest to highlight areas where MCC is not significantly positive using hatching.

Thank you for your advice. In the colorbar, white represents the zero value. Therefore, it is meaningful to plot MCC values not significantly different from zero in white.

Line 183-184: Can the authors explain why the hit rate and false alarm rate are less effective metrics for assessing MHWs?

On the one hand, if we simply classify every day as MHW days, this leads to a high hit rate of 100%. On the other hand, if we simply classify every day as non-MHW days, this leads to a low false alarm rate of 0. However, it is evident that the above two naive classifiers have no skill in detecting MHWs, although they have either high hit rate or low false alarm rate. In contrast, the MCC yields a high score only when predictions perform well in all the four components of the confusion matrix. It has been proved to be an effective metric for imbalanced class classification (Chicco and Jurman 2020).

Line 185: "Further" should be replaced with "Furthermore".

Revised.

Line 203-204: Can the authors clarify the meaning of "standard error"? Is it the range of globally-averaged MCC across 6 datasets?

The standard error (SE) is the standard deviation of sample mean (note that the sample mean is a random variable). It provides an indication of how the sample mean may deviate from the population mean. A smaller SE indicates a more precise estimate. The SE is calculated using the following formula (James et al. 2013):

$$SE = \frac{\sigma}{\sqrt{n}},$$

where σ is the standard deviation of the population and n is the sample size.

In this study, the standard deviation σ of globally averaged MCC for each dataset is estimated as:

$$\sigma = \sqrt{\frac{\sum_{i=1}^{n}(MCC_i - \overline{MCC})^2}{n-1}}$$

where MCC_i is the globally averaged MCC for the i-th dataset, $\overline{MCC}$ is the ensemble (sample) mean over all the datasets and $n = 6$ is the number of datasets (sample size). The standard error is then computed as:

$$SE = \sqrt{\frac{\sum_{i=1}^{n}(MCC_i - \overline{MCC})^2}{n-1}} / \sqrt{n}.$$

Line 279: I am not quite sure if "physics-informed statistical learning method" would be the most accurate way to describe the proposed approach. I think it is a seasonally-varying linear regression model that built on the linear relationship of key ocean dynamics.

Thank you for your comment. We have changed the relevant description to "dynamics-guided statistical learning method". The ocean dynamics suggest a linear but spatio-seasonally varying relationship, guiding the proposal of statistical models.

Line 305: I believe that the following should be the right reference for GLORYS. Lellouche, J., et al. (2021). The Copernicus Global 1/12 Oceanic and Sea Ice GLORYS12 Reanalysis, Frontiers in Earth Science, 9.

Thanks. Revised.

Line 325-326: Why is "T_c" irrelevant to MHWs? Is it because "T_c" was removed from T to get T'? This sentence seems confusing to me.

Yes, subsurface temperature anomaly T' is computed by subtracting the climatological mean seasonal cycle of T (T_c) from T, i.e., $T' = T - T_c$. To avoid confusion, we have revised this part as:

"Note that $T = T' + T_c$ and $\theta = \theta' + T_c$ where T_c is the climatological mean seasonal cycle of T and θ' is the seasonally varying 90th percentile of T'. As MHWs are defined based on $T - \theta$ (Hobday et al. 2016), T_c is irrelevant to MHWs and we can detect subsurface MHWs based on T' alone (Wang et al. 2022)."

Please see Line 423-426.

Line 416: What is the use of the random noise in Eqn (13)?

The random noise is a catch-all for what are missed in the GSVC model. We have added this explanation in the revised manuscript. Please see Line 439-440.

Line 484: Can the authors explain why using a smaller weight on the "one" class and a larger weight on the "zero" class makes the model pay more attention to occurrence of MHW events? I thought a larger weight would mean a larger penalty for missing a prediction of a certain class. In the authors' case, that would mean the model really don't want to predict a wrong "zero".

We are grateful to you for pointing out the typos. We actually use 0.80 for the "one" class (an MHW) and 0.20 for the "zero" class (not an MHW). We have corrected the typo. Please see Line 503-505.

Line 574: I think it is more compelling to plot the figure within the range of 0 to 1.

Thanks for your advice. In this study, the computed false alarm rate is below 0.1. Plotting the false alarm rate with the range of 0 to 1 does not provide clear illustration of its geographical distribution (Fig. R1 and R2).

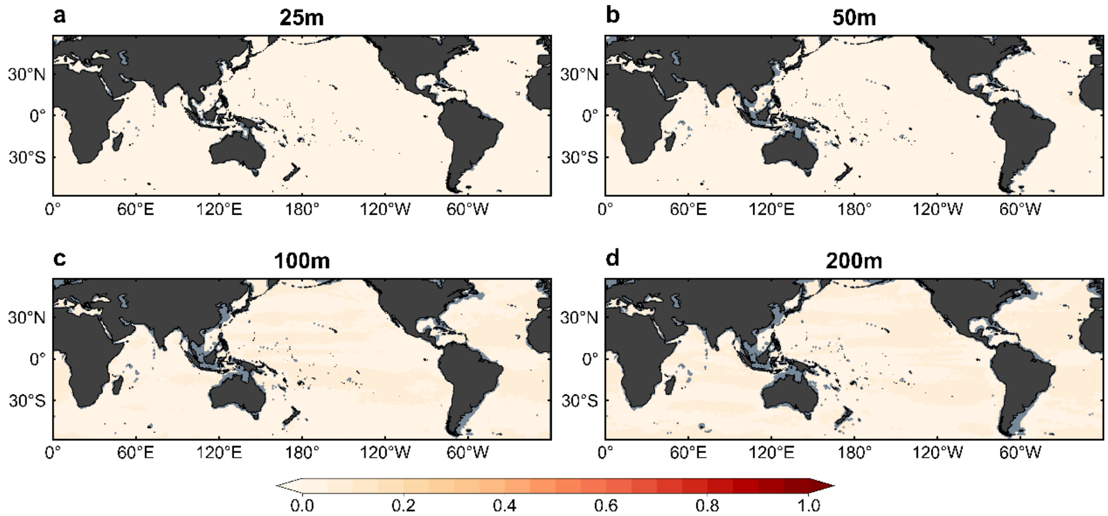

Fig. R1 Global distribution of the false alarm rate for MHW detection at 25 m (a), 50 m (b), 100 m (c) and 200 m (d) averaged over the 6 combinations of training and test datasets, ranging from 0 to 1. Land is marked in grey.

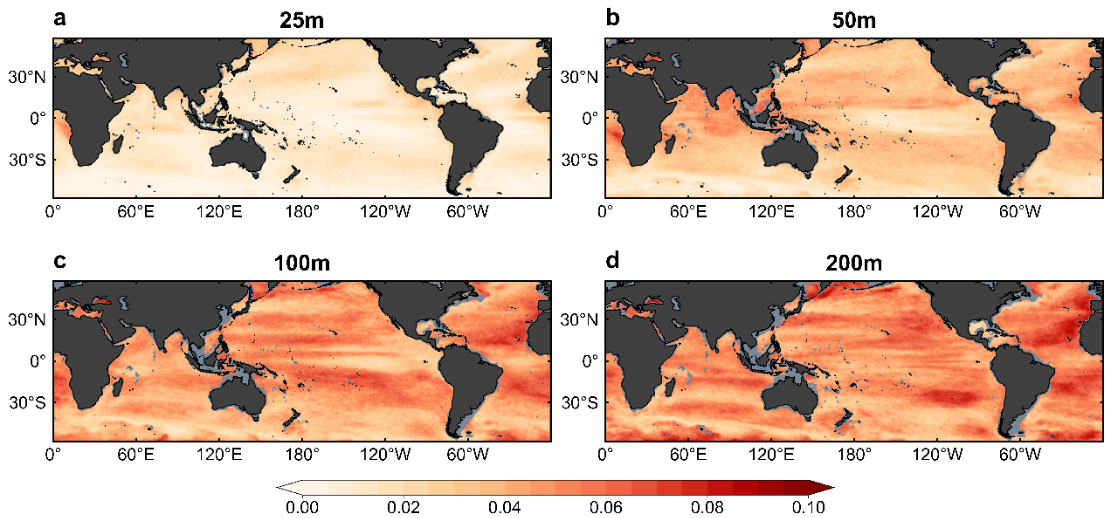

Fig. R2 Global distribution of the false alarm rate for MHW detection at 25 m (a), 50 m (b), 100 m (c) and 200 m (d) averaged over the 6 combinations of training and test datasets, ranging from 0 to 0.1. Land is marked in grey.

Reference list

Chicco, D., and G. Jurman, 2020: The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation. *BMC Genomics*, **21**, 6, https://doi.org/10.1186/s12864-019-6413-7.

Hobday, A. J., and Coauthors, 2016: A hierarchical approach to defining marine heatwaves. *Prog. Oceanogr.*, **141**, 227–238, https://doi.org/10.1016/j.pocean.2015.12.014.

James, G., D. Witten, T. Hastie, and R. Tibshirani, 2013: *An Introduction to Statistical Learning with Applications in R*. Springer,.

Wang, S., Z. Jing, D. Sun, J. Shi, and L. Wu, 2022: A new model for isolating the marine heatwave changes under warming scenarios. *J. Atmos. Ocean Tech.*, **39**, 1353–1366, https://doi.org/10.1175/JTECH-D-21-0142.1.

Reply to the first reviewer

We are very grateful to you for your time in carefully reading our manuscript and providing helpful comments that make our manuscript better. We have carefully considered each of your comments (in blue) and revised the manuscript accordingly. Please find our response (in black) to your comments below.

Reviewer #1 (Comments to Author):

The authors have addressed my previous comments. I find the revised manuscript acceptable for publication.

Thanks. We are grateful to your decision.

Reply to the second reviewer

We are very grateful to you for your time in carefully reading our manuscript and providing helpful comments that make our manuscript better. We have carefully considered your comments (in blue) and revised the manuscript accordingly. Please find our response (in black) to your comments below.

Reviewer #2 (Remarks to the Author):

To support authors' claim that the main contribution of their study that the outperformance of the GSVC over the OLS is meaningful, it should be demonstrated in the Introduction that the common approach in their research field is still based on the simple OLS that universally used for all seasons. I believe that this paper is worth publishing, only demonstrating that the OLS is still widely accepted in their research field.

(Personally, I don't believe their arguments that their findings, that the GSVC is better than the OLS, is valuable in their research fields. It is like a common sense to me, as it is already well recognized that the relationships between physical variables varies seasonally due to their seasonal cycle.)

Thanks for your comment. The GSVC does not only outperform the OLS but also the more complicated method, i.e., the CNN. Both the OLS and CNN assume a temporally constant relationship which is widely accepted in the research field (Guinehut et al. 2012; Su et al. 2018; Meng et al. 2022; Xie et al. 2022). To the best of knowledge, this is the first time a spatially and seasonally varying relationship is adopted to retrieve the subsurface temperature anomaly based on SSHA and SSTA. We have provided more references including the latest references to demonstrate our argument.

As you pointed out, the ocean dynamics suggest a spatially and seasonally varying linear relationship. However, such dynamical insight was not exploited to guide statistical learning before our study. The usage of ocean dynamics in statistical learning is thus a major contribution of our study.

References

Guinehut, S., A.-L. Dhomps, G. Larnicol, and P.-Y. Le Traon, 2012: High resolution 3-

D temperature and salinity fields derived from in situ and satellite observations. *Ocean Sci.*, **8**, 845–857, https://doi.org/10.5194/os-8-845-2012.

Meng, L., C. Yan, W. Zhuang, W. Zhang, X. Geng, and X.-H. Yan, 2022: Reconstructing High-Resolution Ocean Subsurface and Interior Temperature and Salinity Anomalies From Satellite Observations. *IEEE Trans. Geosci. Remote Sensing*, **60**, 1–14, https://doi.org/10.1109/TGRS.2021.3109979.

Su, H., L. Huang, W. Li, X. Yang, and X. Yan, 2018: Retrieving Ocean Subsurface Temperature Using a Satellite-Based Geographically Weighted Regression Model. *J. Geophys. Res. Oceans*, **123**, 5180–5193, https://doi.org/10.1029/2018JC014246.

Xie, H., Q. Xu, Y. Cheng, X. Yin, and Y. Jia, 2022: Reconstruction of Subsurface Temperature Field in the South China Sea From Satellite Observations Based on an Attention U-Net Model. *IEEE Trans. Geosci. Remote Sensing*, **60**, 1–19, https://doi.org/10.1109/TGRS.2022.3200545.

Reply to the third reviewer

We are very grateful to you for your time in carefully reading our manuscript and providing helpful comments that make our manuscript better. We have carefully considered each of your comments (in blue) and revised the manuscript accordingly. Please find our response (in black) to your comments below.

Reviewer #3 (Comments to Author):

The authors have largely addressed my comments. And I really appreciate the effort. I have only a few very minor comments left. The first and foremost important one is that I think the equations should be moved back to the method section. I don’t think simply moving the equations from the method section to the result section is going to help readers understand the methodology. I think both the first reviewer and I were asking for physical description of the method.

Thanks for your advice. We have moved the equations back to the method section and only retained the dynamical description in the main text.

Other minor comments are as followed:

Line 53: I have no expertise in biology. But it seems weird to categorize salmons as “large organisms”. Perhaps just use “fish” is sufficient.

Thanks for your advice. We have revised it. Please see Line 53.

Line 215-223: Perhaps the authors should explicitly mention how they treat climate change. Before training the model, do the authors detrend first so the system represents a stationary system?

Thanks for your question. All the models (OLS, CNN GSVC) are trained based on the undetrended data. We do not detrend the data, because it is a common practice to define MHWs based on undetrended temperature (Hobday et al. 2016; Oliver et al. 2018).

Line 252-253: If I am not color blind, from my view, the land is in dark grey color, whereas insignificant areas are in light grey color?

Sorry for the unclarity. The land is marked in dark grey and the continental margins are marked in light grey.

Line 296-298: I wouldn't say that OLS regression and CNN do not "exploit ocean dynamics". They are still trying to capture the relationship between surface and subsurface ocean. They get reasonable results due to the fact that they were able to capture some of that relationship.

Sorry for the imprecise expression. We have revised the description to "do not effectively exploit ocean dynamics" for OLS regression and CNN.

Line 359-365: I think that a fundamental issue is the lack of the long record for use in the more complicated methods. Perhaps the authors could comment on that the paragraph.

Thanks for your comment. Please see Line 247-248 in the revised manuscript.

References

Hobday, A. J., and Coauthors, 2016: A hierarchical approach to defining marine heatwaves. *Prog. Oceanogr.*, **141**, 227–238, https://doi.org/10.1016/j.pocean.2015.12.014.

Oliver, E. C. J., and Coauthors, 2018: Longer and more frequent marine heatwaves over the past century. *Nat. Commun.*, **9**, 1324, https://doi.org/10.1038/s41467-018-03732-9.