

Authors' Response to Referee #2

We sincerely thank the referee for the thorough and constructive assessment of our manuscript. We recognize the concerns regarding the conceptual coherence of the manuscript, the relationship between continuous SST forecasting and subsequent MHW identification, the use of “physics-informed” terminology, and the adequacy of the forecast verification. In response, we have checked the SEDI calculation underlying Fig. 14 against its original definition. We will correct the inconsistent metric definitions, distinguish daily threshold-exceedance detection from duration-based MHW identification, and strengthen the corresponding verification. We will also address the comments concerning simple forecasting baselines, the multi-lead prediction setup, and the alignment between the stated MHW metrics and the reported evaluations. Below, the referee’s comments are reproduced in black, followed by our point-by-point responses in blue.

Summary and General Comments

Referee comment

This manuscript assesses the prediction of SST extremes/MHWs using SwinTrans-ConvLSTM. In terms of datasets the authors are using OSTIA SST and ERA5 to represent the associated physical parameters and the study area is East China Sea. U and V winds and ΔT (SST – 2m air temperature) are used as dynamical and thermodynamical physical constrains. Definition and metrics from Hobday et al., 2016, Oliver et al., 2018 and Xu et al., 2025 are used to characterize MHWs. Four metrics were introduced to characterize MHWs: mean duration, mean intensity, occurrence frequency, and total MHW days. Metrics like MAE/RMSE/ R^2 /PCC account for SST field prediction whereas SEDI/POD/FAR measure the discrete extreme even detection skill. The authors deploy a set of deep learning (DL) models to build a comparison between them and assess the best detection skill of all: LSTM, ConvLSTM,

Transformer, Swin Transformer and SwinTrans-ConvLSTM. They conclude that the hybrid DL framework, SwinTrans-ConvLSTM is best to predict SST extremes with lead-times up to 7 days with high accuracy.

Therefore, the manuscript presents an extensive set of experiments, figures and metrics, but lacks a consistent conceptual line. Specifically, the relationship between SST/SST anomalies and MHW event prediction is never clearly established and shifts across sections and metrics. I'm finding it difficult to exactly establish the manuscript's contribution lane. Combined with the technical issues described below (written as suggested revisions, I recommend rejection, rather than revision, since it'll need a full restructuring instead of incremental fixes.

Response

We appreciate the referee's constructive assessment and acknowledge that the relationship between continuous SST forecasting, daily SST threshold-exceedance detection, and MHW event identification was not sufficiently clearly established in the original manuscript. The primary contribution of our study is the development and evaluation of the hybrid SwinTrans-ConvLSTM framework for 1–7-day SST forecasting at 0.05° resolution in the East China Sea. The subsequent extreme-event-related analysis evaluates daily exceedances of a seasonally varying SST threshold derived from the predicted SST, rather than the forecasting skill of complete duration-based MHW events. We will clarify these distinct evaluation targets throughout the manuscript and avoid interpreting improvements in continuous SST prediction accuracy as direct evidence of improved MHW event prediction skill.

We have inspected the script used to generate Fig. 14 and confirmed that its SEDI calculation follows the hit-rate and false-alarm-rate definitions of Ferro and Stephenson (2011), with a small pseudocount added for numerical stability. In particular, the false-alarm rate uses $FP/(FP+TN)$, rather than the false-alarm ratio $FP/(FP+TP)$ stated in the

manuscript. We will correct this inconsistency, explicitly define the contingency-table entries, and cite the original SEDI reference.

The same inspection shows that Fig. 14 evaluates daily SST threshold exceedances of the seasonally varying 90th-percentile threshold without applying a minimum-duration criterion. Following the referee’s option in Major Comment 5, we will state this scope explicitly throughout the manuscript, clearly distinguish daily threshold-exceedance skill from continuous SST and SST-anomaly evaluations, and limit our conclusions to the quantities that are actually evaluated. We will explicitly distinguish this daily threshold-exceedance analysis from duration-based MHW event identification, revise the associated figure captions, metric descriptions, and interpretation accordingly, and avoid presenting daily threshold-exceedance skill as evidence of complete MHW event forecasting skill.

In the revised manuscript, we will align the metrics introduced in Sect. 2.4 and Table 1 with the evaluations reported in the Results and ensure that each metric is associated with a clearly defined prediction or detection target. We will also replace the overstated “physics-informed” terminology with “multivariate input features” and the gradient term as “spatial-gradient regularization”, include a simple SST persistence baseline, and align Eq. (2) and its explanation with the implemented direct multi-lead forecasting strategy. The figure captions and discussion of the regional framework’s limitations will be expanded accordingly. We earnestly request the opportunity to revise our manuscript.

Major Comments:

1. Inflated physics-informed terminology throughout the manuscript. I agree on the MHW rarity and the need of an extra processing step to balance the classes, and L_{grad} used as spatial/sharpness smoothness here is a good choice. However, toning down the language and naming of the method regarding the physics-informed concept is necessary.

Response

We thank the referee for this constructive comment. We agree that the term “physics-informed” overstates the physical constraints imposed by the present framework. Neither the network architecture nor the training objective explicitly enforces governing physical equations or conservation laws. The wind components (u_{10} and v_{10}) and air–sea temperature contrast (ΔT) are incorporated as predictors selected for their physical relevance. We will therefore describe them as “multivariate input features” and clarify that their inclusion does not impose the associated dynamical or thermodynamical relationships on the network.

We also agree that L_{grad} is a mathematical regularization term for spatial structure rather than a physical conservation constraint. In the implementation, it penalizes the L1 discrepancy between the magnitudes of adjacent-grid finite differences in the predicted and reference fields along the two horizontal directions. This term is intended to encourage agreement in local spatial variation and discourage excessive smoothing. We will consistently refer to it as “spatial-gradient regularization”. The Huber loss, seasonal weighting, and staged online hard example mining (OHEM) will likewise be described as components of the optimization strategy, rather than as physical constraints.

Accordingly, we will revise the terminology and associated interpretations throughout the entire article. Specifically, Sect. 2.3 will be renamed “Multivariate Input Features and Data Preprocessing”; the training objective in Sect. 2.6.1 will be described as a “Gradient-Regularized Curriculum Loss”. and Sect. 4.1 will be renamed “Evaluation of Multivariate Input Features”. These revisions will distinguish the physical motivation for selecting predictors from the mathematical role of regularization and avoid presenting predictive improvements as evidence that the network enforces physical laws or resolves particular physical mechanisms.

2. SEDI equation seems to have a miscalculation: In the original paper that defined SEDI (Ferro and Stephenson, 2011), F is defined as $FP/FP+TN$, representing the rate out of actual non-occurrences, how many time did we guess wrong, named ‘false alarm rate’. In the cited paper (Shu et al 2025), F is defined as $FP/FP+TP$ and named ‘false alarm rate’. In the reviewed manuscript, the authors call FAR (F) as ‘false alarm ratio’ (and later on rate) and define it as $FP/FP+TP$. Is the substitution of TN with TP intentional in the manuscript – simply following Shu et al.? In Shu et al there is no documented reasoning for this substitution either. Do the authors consider something different than what the original Ferro and Stephenson calculation considers? Could the authors please double check the sources, their definitions and goals with the metric? Also, the explanation of "Hits" denote correctly predicted marine heatwave days, "Misses" represent missed observations, and "False Alarms" correspond to false detections ’ might be misleading (to me as well), could the authors possibly add the traditional TP, TN, FP and FN naming to make this calculation clearer? Where: TP = True Positive, TN = True Negative, FP = False Positive, and FN = False Negative. Full reference of original paper: Ferro, C. A. T., and D. B. Stephenson, 2011: Extremal Dependence Indices: Improved Verification Measures for Deterministic Forecasts of Rare Binary Events. *Wea. Forecasting*, **26**, 699–713, <https://doi.org/10.1175/WAF-D-10-05030.1>.

Response

We thank the referee for identifying this inconsistency and for directing us to the original definition of SEDI. We agree that the manuscript conflates the false-alarm ratio with the false-alarm rate. Equation (15), expressed using the conventional contingency-table notation, defines the false-alarm ratio as $FP/(TP + FP)$. Although this is a valid verification metric, it is not the false-alarm rate required in the SEDI formulation of Ferro and Stephenson (2011). Its use as the definition of the quantity entering Eq. (13) is therefore incorrect. We did not intend to introduce a modified SEDI or a different interpretation of this index.

Following the original definition, we will explicitly distinguish the hit rate (H), equivalent to the probability of detection (POD), from the false-alarm rate (F). These quantities will be defined as:

$$H = \text{POD} = \frac{TP}{TP + FN}, \quad F = \frac{FP}{FP + TN}$$

The SEDI formulation will be presented as

$$\text{SEDI} = \frac{\ln F - \ln H + \ln(1 - H) - \ln(1 - F)}{\ln F + \ln H + \ln(1 - H) + \ln(1 - F)}$$

We will also introduce the following contingency table to make the definitions unambiguous:

Forecast classification	Reference positive	Reference negative
Positive	True positive (TP; hit)	False positive (FP; false alarm)
Negative	False negative (FN; miss)	True negative (TN; correct negative)

Importantly, our inspection of the script used to generate Fig. 14 shows that the implemented calculation uses $FP/(FP + TN)$. Thus, the script uses the false-alarm rate required by SEDI, rather than the false-alarm ratio written in Eq. (15) in the original manuscript. The implementation also adds a small pseudocount of (10^{-6}) to each of the four contingency-table counts before calculating the rates to avoid undefined logarithms at the boundaries. We will document this numerical treatment explicitly and specify the pseudocount value used in the implementation.

We will revise Eqs. (13)-(15) and the accompanying text to use consistent notation and clearly distinguish the false-alarm rate from the false-alarm ratio. We will cite Ferro and Stephenson (2011) directly as the methodological source for SEDI. We will also clarify the classification underlying Fig. 14: the current script evaluates daily SST exceedance of the seasonally varying 90th-percentile threshold. This daily exceedance classification will be distinguished explicitly from MHW event identification that

additionally applies a minimum-duration criterion (Hobday et al., 2016). These revisions will make both the calculation and the scope of its interpretation transparent.

3. I understand that the paper compares four deep learning benchmark models, however a simple SST persistence model to use as a baseline would be very valuable to make the deep learning choice more solid. At minimum a naive persistence $Y(t+n) = X(t)$ without parameters and training – to represent a baseline for the highly autocorrelated SST. Another suggestion would be the anomaly persistence: $Y(t+n) = \text{Climatology}(t+n) + (\text{SST}(t) - \text{Climatology}(t))$ – should also be straight forward given that daily climatology has already been calculated in the manuscript. An AR(1) baseline fit per gridpoint on the SST anomaly could be another option. Any one of the above would be sufficient as a baseline model and fulfill this review point. These results would not need to be presented in the main manuscript (could fit well in a supplementary information), but the skill comparison between the simple baseline and DL models should be seen in the text (somewhere in the results section around table 3).

Response

We thank the referee for this constructive suggestion. We agree that simple forecasting baselines are necessary to assess the predictive value of the deep learning models relative to the strong temporal persistence of SST. Following this recommendation, we have now completed two additional baseline experiments: a parameter-free SST persistence forecast and a gridpoint-specific first-order autoregressive model fitted to SST anomalies [AR(1)].

For the persistence baseline, the SST at the forecast origin (t) is retained throughout the seven-day forecast:

$$\hat{T}_{\text{pers}}(t + n | t) = T(t), \quad n = 1, \dots, 7.$$

For the AR(1) baseline, daily SST anomalies are defined relative to a calendar-day climatology calculated exclusively from the training period (1982–2015), with a 31-

day circular smoothing window applied to the climatological annual cycle. At each ocean grid point, an intercept and an autoregressive coefficient are estimated from consecutive training-period anomaly pairs using least squares, with the autoregressive coefficient constrained to the interval $[-0.999, 0.999]$. Starting from the SST anomaly at the forecast origin, the model recursively predicts anomalies for leads 1–7 and then adds the training-period climatology for each corresponding target date:

$$\hat{a}(t + n|t) = \alpha + \phi \hat{a}(t + n - 1|t), \quad n = 1, \dots, 7$$

$$\hat{T}_{AR1}(t + n|t) = C(t + n) + \hat{a}(t + n|t)$$

where C denotes the training-period climatology, α and ϕ are the gridpoint-specific intercept and autoregressive coefficient, and the initial anomaly is

$$\hat{a}(t | t) = T(t) - C(t).$$

Neither the climatology nor the AR(1) coefficients were fitted or tuned using validation or test data. Both baselines use the observed SST at the forecast origin for initialization; for AR(1), this SST is converted to an anomaly relative to the training-period climatology. No subsequent observed SST is used during the seven-day forecast.

The evaluation covered 1,807 forecast origins from 14 January 2020 to 24 December 2024, seven forecast leads, and 7,289 ocean grid points. Both baselines were evaluated against the same processed SST reference fields used to evaluate the original model forecasts, using matched forecast cases, spatial masks, and validity criteria. Each lead contained 13,171,223 valid grid point–forecast pairs.

As shown in Table R1, the persistence baseline yielded a pooled MAE of 0.5909 °C and RMSE of 0.7694 °C across leads 1–7. The anomaly AR(1) baseline yielded a pooled MAE of 0.4420 °C and RMSE of 0.5790 °C, with lower MAE and RMSE than persistence at every lead. By comparison, the previously reported pooled RMSE of SwinTrans-ConvLSTM across leads 1–7 is 0.1555 °C, approximately 73.1% lower than

the AR(1) baseline RMSE of 0.5790 °C. This comparison provides a quantitative assessment of the deep learning model relative to both a parameter-free persistence benchmark and a grid point-specific linear autoregressive benchmark.

In the revised manuscript, we will include the baseline formulations and experimental settings in the Methods and provide the detailed baseline results in the Supplement. As requested, the comparison with the deep learning models will also be discussed in the main Results section alongside Table 3, using matched reference fields, forecast cases, spatial masks, and metric aggregation.

Table R1. MAE and RMSE of the persistence and anomaly AR(1) SST forecast baselines at leads 1–7.

Forecast lead (days)	Persistence MAE (°C)	Persistence RMSE (°C)	Anomaly AR(1) MAE (°C)	Anomaly AR(1) RMSE (°C)
1	0.4065	0.5429	0.3783	0.5068
2	0.4435	0.5832	0.3819	0.5058
3	0.5000	0.6466	0.4000	0.5241
4	0.5706	0.7269	0.4285	0.5564
5	0.6507	0.8188	0.4633	0.5976
6	0.7371	0.9186	0.5014	0.6435
7	0.8276	1.0236	0.5409	0.6915
Pooled, leads 1–7	0.5909	0.7694	0.4420	0.5790

Notes: The evaluation covered 1,807 forecast origins and 7,289 ocean grid points, yielding 13,171,223 gridpoint–forecast pairs per lead and 92,198,561 pairs across leads 1–7. Pooled MAE and RMSE were calculated from all evaluated errors across the seven leads, with pooled RMSE defined as the square root of the pooled mean squared error.

The baseline formulations and evaluation settings will be described in the Methods, and the lead-specific and pooled results will be provided in the Supplementary Material. As requested, we will also discuss the comparison with the persistence and anomaly AR(1) baselines in Sect. 3.1 alongside Table 3. This comparison will help contextualize the performance of the deep learning models relative to simple forecasting benchmarks.

4. Eq. 2 is missing the multiple lead times component. Given that $Y_{t+n} = X_t + F(X_{t-k:t})$, I'd consider that the DL model predicts all lead times at once – still independently but using the same starting point t . However, looking at Fig 3, a second option may have been considered which every predicted day is calculated based on the one before it so the Eq. 2 should incorporate the lead times such as that: with the base case. In the second option the moving 14-day window would switch one day ahead for every predicted day. The interpretation of Fig 3 and Table 4 would also be affected depending on which of the above two options is valid. Please clarify which one of the two the model implements and revise Eq 2, Fig 3 and table 4 explanations if necessary.

Response

We thank the referee for highlighting the ambiguity in Eq. (2) and its implications for interpreting the forecast results. Our SwinTrans-ConvLSTM model uses a direct, joint multi-output forecasting strategy. For a forecast initialized on day t , the input consists of 14 historical feature frames indexed from $t-13$ to t , and the network produces seven output fields corresponding to days $t+1$ through $t+7$. These outputs share the same input

sequence and learned representations, with a separate output channel for each lead time. A predicted SST field is not fed back into the model to generate the next lead time.

To make the multiple lead times and their common forecast origin explicit, we will revise Eq. (2) to express the complete forecasting procedure as

$$\hat{Y}(t + n | t) = [G_\theta(Z_{t-13:t})]_n, \quad n = 1, \dots, 7$$

where $Z_{t-13:t}$ denotes the 14-frame multivariate input sequence, $\hat{Y}(t + n | t)$ denotes the SST field predicted for day $t + n$ from forecast origin t , and G_θ represents the complete forecasting mapping, including the network and the conversion of its outputs to SST units. The mapping returns seven SST forecast fields jointly, and $[\cdot]_n$ selects the field corresponding to lead n . Specifically, the network predicts standardized SST anomalies, which are transformed back to temperature units and combined with the training-period climatology for the corresponding target dates. We will revise the explanation accompanying Eq. (2) to describe this prediction and reconstruction procedure accurately. We will clarify in Eq. (2) and its accompanying text that the equation represents the seven outputs jointly, while the anomaly-to-SST reconstruction is applied to the corresponding output fields.

The 14-day input window remains fixed while all seven lead-time outputs are generated for a given forecast origin. It advances by one day only when a new forecast is initialized at the next origin. Thus, successive forecast instances use moving historical windows, but the seven predictions within an individual instance are not generated by recursively advancing the input window with predicted values.

We will also revise the captions and accompanying explanations of Fig. 3 and Table 4 to make this setup explicit. For the proposed model, the error at lead n is calculated by comparing the direct forecast $\hat{Y}(t + n | t)$ with the reference SST for day $t + n$, aggregated over the evaluated forecast origins and valid ocean grid points. The variation in error across lead times therefore describes the performance of the corresponding

direct outputs; it should not be interpreted as error accumulation caused by recursively feeding earlier predictions back into the model.

5. MHW metrics evaluation missing. Section 2.4 and Table 1 describe 4 metrics on which the model would be evaluated upon predicting MHWs following Hobday et al., 2016 and Xu et al., 2025 definitions. These are: mean duration, mean intensity, frequency and total MHW days. However, the Results section 3.5 presents only the intensity (Figs 10-13) as a predicted vs observed comparison. The rest of the metrics are not shown at all. This creates some confusion throughout the manuscript, as to whether we are assessing the skill of predicting MHW events or SST extremes as a continuous field and adding the MHWs characterization on top. I'd recommend either (a) extending the results to show the rest of the MHWs metrics comparisons between predicted and observed occurrences, or (b) if that's out of scope, clarify in the text that the manuscript's primary contribution is detection of SST/SST anomaly fields with an additional binary MHW-detection skill (SEDI), and reframe results' section 3.5 accordingly.

Response

We thank the referee for identifying the need to align the metrics introduced in Sect. 2.4 and Table 1 with the results presented in Sect. 3.5. We agree that the manuscript should clearly distinguish general MHW characteristics from the quantities actually used to evaluate forecasts. We will follow the referee's option **(b)** and clarify the scope of the study accordingly.

The primary contribution of this study is the development and evaluation of the SwinTrans-ConvLSTM framework for 1–7-day SST forecasting in the East China Sea, supplemented by analyses of SST anomalies and threshold exceedances relevant to marine heatwaves. We will reorganize the presentation around three distinct aspects: continuous SST forecast accuracy, the evaluated SST-anomaly statistics, and daily threshold-exceedance skill. This structure will make the relationship between SST

forecasting and the subsequent analysis of extreme warm conditions explicit. In particular, we will distinguish the characterization of MHWs from the verification of forecast skill for complete MHW events.

Specifically, we will revise Sect. 2.4 and reorganize Table 1 to present the definitions of the quantities used in the reported verification. Duration, occurrence frequency, and total MHW days will be removed from the stated set of evaluated forecast metrics. We will also reframe Sect. 3.5 as “Evaluation of Extreme SST Conditions” and clarify the definitions, sample selection, and averaging procedures underlying Figs. 10–13. We will explicitly identify the quantity shown in these figures and distinguish SST-anomaly statistics calculated for threshold-exceeding days from the conventional event-based MHW intensity metric. Where SST anomalies are averaged over threshold-exceeding days within a seven-day forecast window, this calculation will be stated explicitly in the text and figure captions.

Consistent with our response to Major Comment 2, we will clarify that Fig. 14 evaluates daily SST exceedance of the seasonally varying 90th-percentile threshold. The corresponding binary skill scores will be interpreted specifically in relation to this daily classification and discussed separately from the continuous SST and SST-anomaly evaluations. We will not interpret these daily threshold-exceedance scores as direct verification of MHW events satisfying the minimum-duration criterion.

Finally, we will align the Abstract, Discussion, and Conclusions with this scope. Our forecast-performance conclusions will focus on SST accuracy, the evaluated SST-anomaly statistics, and daily threshold-exceedance skill. Event duration, occurrence frequency, and total MHW days will not be presented as quantitatively verified forecast outcomes. We will also avoid inferring improved MHW event prediction solely from improvements in aggregate SST accuracy.

Minor Comments:

1. Please add a few sentences on the potential drawbacks of the selected ML models combination. E.g. the combination is more expensive, or it is hard to generalize on a global scale vs a regional study.

Response

We thank the referee for this suggestion. We will add a paragraph to the Discussion addressing the limitations of the hybrid framework. Combining the ConvLSTM temporal encoder with Swin Transformer blocks increases architectural complexity because both components must be configured and tuned. We will also discuss potentially substantial computational and memory requirements associated with processing high-resolution spatiotemporal inputs. Applying the framework to a substantially larger spatial domain may further increase these demands and require adjustments to the data-processing pipeline and training strategy.

Furthermore, the present evaluation is regional, and the results do not establish that the framework would achieve comparable performance in other ocean basins or at the global scale. Differences in oceanographic conditions, seasonal variability, and data characteristics may affect transferability. We will therefore clarify that applications beyond the study region require independent validation and may necessitate model adaptation or retraining using representative data from the target region.

2. Add stand-alone figure captions. Many figure captions are just a short line of naming the panels. I'd argue that figure captions need to be stand-alone descriptions of the figures thinking that readers that would just want to skim through the paper could understand/interpret in isolation. Please expand the captions to be more self-contained.

Response

We agree that the figure captions should allow readers to understand the figures without repeatedly consulting the main text. We will revise the captions throughout the manuscript to make them self-contained by identifying the variables or metrics shown, the meaning of each panel, the relevant dates or evaluation period, forecast lead times, units, and the meanings of colors, contours, symbols, and reference lines, as applicable. Where applicable, we will also specify the reference data, sample-selection criteria, and spatial or temporal averaging used to construct the plotted quantities.

In particular, the captions of the lead-time comparisons will clarify that the proposed model produces direct multi-lead forecasts from a common forecast origin. The captions of Figs. 10–13 will specify the SST-anomaly quantities and averaging periods represented, including the distinction between quantities evaluated at individual forecast lead times and statistics aggregated across forecast lead days 1–7, where applicable.

The caption of Fig. 14 will identify the daily threshold-exceedance classification and explain the basis on which the reported skill scores are calculated and summarized. These changes will be consistent with the methodological clarifications described in our responses to the Major Comments.

3. Grammar/language polishing throughout (e.g. ‘this study adopt’ in line 132; ‘As seasons progresses’ in line 385; etc).

Response

We thank the referee for pointing out these language issues. We will review the entire manuscript for grammar, sentence structure, and consistency of terminology. The examples identified will be corrected to “this study adopts” and “As the seasons progress”, respectively. We will also check subject–verb agreement, article usage, singular and plural forms, and verb tense throughout the text and captions. Terminology

related to the model inputs, spatial-gradient regularization, forecast lead times, and verification metrics will be standardized to ensure consistency with the methodological clarifications described in our responses to the other comments.

We thank the referee again for the careful assessment and constructive suggestions. We will revise the manuscript accordingly, with particular attention to clarifying the study's scope, making the forecasting and verification procedures explicit, and ensuring that the conclusions are supported by the results presented.

References

Ferro, C. A. T., and Stephenson, D. B.: Extremal Dependence Indices: Improved Verification Measures for Deterministic Forecasts of Rare Binary Events, *Weather and Forecasting*, **26**, 699–713, <https://doi.org/10.1175/WAF-D-10-05030.1>, 2011.

Hobday, A. J., Alexander, L. V., Perkins, S. E., Smale, D. A., Straub, S. C., Oliver, E. C. J., Benthuisen, J. A., Burrows, M. T., Donat, M. G., Feng, M., Holbrook, N. J., Moore, P. J., Scannell, H. A., Sen Gupta, A., and Wernberg, T.: A hierarchical approach to defining marine heatwaves, *Progress in Oceanography*, **141**, 227–238, <https://doi.org/10.1016/j.pocean.2015.12.014>, 2016.