Responses to comments of Anonymous Referee #2

We want to thank reviewer #2 for providing helpful comments on our manuscript. Below we present the reviewer's comments in black, while our responses are in blue and directly follow each comment. Note that we have separated the comments into "general comments" and "specific comments", where the latter are the comments which were added directly to the manuscript.

General comments

The authors have reduced the observation error variance when supermodding by a factor of more than six. While some change in observation error would be expected, no proper argument is given for the values chosen. From their results I think it is possible that this was too large a reduction and the authors are overfitting the data. Also, I would have expected to see an additional experimental run where the observation errors have not been changed so the effect of the supermod operator could have been seen in isolation.

Alpha (α) was chosen for each of the satellite products (IR, PMW, and PMW with supermod) based on a number of experiments with different choices of α. The results were validated against SLSTR Sentinel-3A SSTs. These experiments did not show an increase in the errors for the low value of α = 0.3 compared to the other values tested. However, we have now reassessed the observation errors used, following the diagnosis proposed in Desroziers et al. (2005), and found that this value is indeed too low. Increasing α does, however, not improve the error statistics. The spectrum on the other hand shows that there were structures present in the SST fields probably caused by this overfitting, and this was particularly visible for cycles containing observations with undetected radio frequency interference (RFI). We have thus re-run the experiments with the supermod operator activated (PMW2 and COMB2) with the same value for α (α = 2) as in PMW1 and COMB1. Since we saw that the results were affected by bad observations, we also removed PMW observations within 75 km of oil rig locations in our domain. To ensure a fair comparison of PMW2/COMB2 and PMW1/COMB1 we also re-ran the latter two experiments without PMW observations in the same regions. The manuscript has thus been revised in accordance with the new results.

The results shown indicate an increase in power in the high wavenumbers in the model. However, I don't understand this as the increments are visually smoother

(figure 7). This needs to be properly explained. Also, the power spectrum of the increments, not just the model, should be given.

The increments are indeed smoother when the supermod operator is activated. This is demonstrated in Sect. 3.1, where increments for varying footprint sizes are compared. That the more detailed increments of PMW1 yield smoother SST fields can be explained by the fact that these detailed increments often reflect the removal or dampening of small-scale features present in the model background SST fields. The smoother increments of PMW2 act more as an adjustment of the mean of a larger area. To illustrate this, plots showing the increment, background, and analysis fields for an assimilation cycle for PMW1 and PMW2 are shown below (for an area of 200 km x 200 km). The PMW experiments show enhanced heating in regions with cold water in the background, and the areas with abrupt transitions from heating to cooling (e.g. between 70–96 km on the y-axis, 0–24 km on the x-axis) are associated with fronts in the background that are weakened or even removed completely in the analysis. In PMW2, the structures found in the background are only slightly adjusted, but the analysis is overall warmer than the background. In the updated manuscript we now explain that highly detailed increments do not necessarily result in more structures in the analysis, as this indeed can seem a bit contradictory. We are, however, not convinced that adding the power spectra of the increments to the manuscript is necessary.

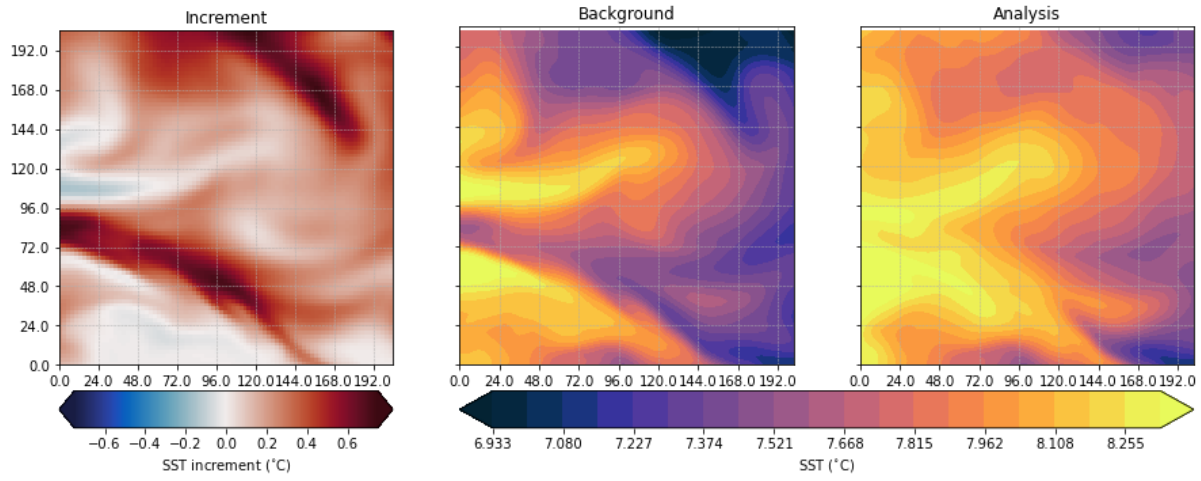

Figure 1: Increment, background SST and analysis SST of PMW1 on 1 June 2018.

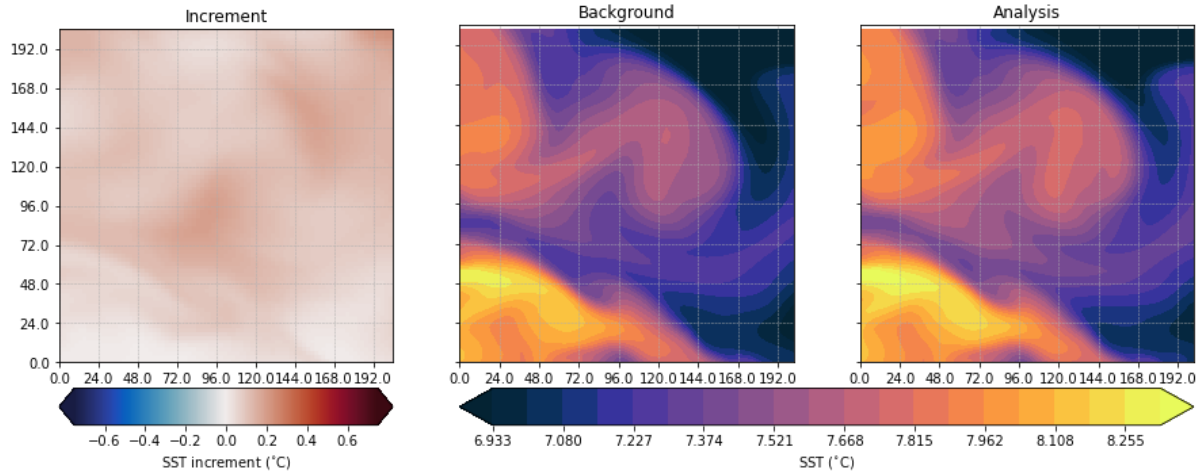

Figure 2: Increment, background SST and analysis SST of PMW2 on 1 June 2018.

The supermoded observations resulted in significantly degraded statistics. I found the authors assertion that this was due to a "double penalty" issue to be unconvincing. This is because neither the resolution of the model nor the verification data has changed. It's also because the increments look smoother to me (see above bullit point). I think it is more likely that they are overfitting the data and putting noise into the model. The authors either need to provide much better evidence to support their beliefs and refute mine, or they need to recalibrate the assimilation to reduce overfitting.

As explained above, we did indeed overfit to the PMW observations in the submitted manuscript. As a consequence, we have re-run the experiments such that we now use the same specification of observation error in all experiments. This is done to ensure that the experiments are comparable. We also conducted the validation using different spatial filters, and in the new set of experiments, there is no evidence to support our suggestion that double penalty could be a part of the explanation. We have thus removed this from the updated manuscript.

Specific comments

P2/L23: Some representative numbers of the resolutions of the IR, PMW, and models, would be useful here.

This is a good idea. Some representative numbers have been added to the manuscript.

P2/L30: *"However, an attempt with regards to assimilating PMW SSTs into a high-resolution regional ocean forecast model has so far not been undertaken."*

This is not true. PMW data from AMSR2 are assimilated into the AMM7 and AMM15 models of the north west european shelf. see https://doi.org/10.1016/j.ocemod.2018.07.004 and https://doi.org/10.5194/os-17-1791-2021.

Thank you for pointing this out. We see now that this sentence does not convey the intended meaning. PMW SSTs are indeed assimilated into regional ocean forecast models, but what we intended to say is that as far as we know, a supermod observation operator (or footprint operator) has not been described and implemented for assimilating PMW SSTs into a high-resolution regional ocean forecast model.

P2/L32: *"called the supermod operator"*
A similar idea is used for sea surface salinity in https://rmets.onlinelibrary.wiley.com/doi/10.1002/qj.3461 and SST in https://www.sciencedirect.com/science/article/pii/S0034425711002197?via%3Dihub. These should be referenced.

These references were already cited in the manuscript (in the previous paragraph). We have reformulated these paragraphs to better emphasize that the references apply similar ideas for a supermod operator.

P2/L41: *"spatial scales of the SST structures typically are small at high latitudes."*
You should justify this statement. I assume you are thinking of the shorter Rossby radius.

Indeed, it is the Rossby radius we were thinking of. We have modified this statement in the updated manuscript.

P4/L75: x and y are vectors, so the usual convention is that they are bold and not italic. Please change - both in the equations and in the main text.

For the mathematical symbols, we followed the submission guidelines which states that vectors should be identified in boldface italics. We have thus not changed the notation.

P4/L78: *"we gather the observations using a temporal spacing of 15 minutes"*
There is insufficient detail to know what this means. Please clarify. My guess is that you assume all obs within each 15 minute block are treated as if they occur at the same time. But this is not the only interpretation. You could be doing some sort of averaging. Please be precise.

We modified the statement so it is more clear that the individual observation times are rounded to the nearest ¼ hour. This sentence is also moved to Sect. 2.2.

P4/L79: *"The observation operator, H, returns the model equivalents of the observations by interpolating the model values to the observation locations"*
What about time interpolation?

We have added a sentence to the manuscript emphasizing the fact that in the case of 4D-Var H includes the nonlinear model which facilitates the temporal mapping.

P4/L83: *"as the 4D-Var minimization algorithm involves the inverse of R"*
You can invert a non-diagonal R, it's just harder.

Yes, indeed. We have changed this part of the paragraph to:
"**R** is assumed to be diagonal in most operational data assimilation systems, including the implementation in ROMS used in this study (Gürol et al., 2014). This assumption means that the observation errors are assumed to be uncorrelated in both time and space."

P4/L97: *"α: α = 0.3 when the supermod operator (see Sect. 3) is activated, and α = 2 when the supermod operator is not activated. "*
Setting alpha = 0.3 for supermod experiments seems arbitrary and difficult to justify to me. There are two reasons for this.

1. It means any comparisons between the supermod and non-supermod experiments are difficult to interpret. Are the changes because you have averaged the model, or because you have reduced the observation error?

2. Supermod is an averaging operator. Its effect on the error should be mathematically predictable to some extent, and related to the number of grid points that have been averaged. I would expect some sort of relation to (1/N) for the variance.

I think this is a really serious issue with what is presented here, and needs to be addressed. At the very least:

A scientifically reasonable justification needs to be made for the use of alpha=0.3, and how it relates to the expected effect of supermod on the actual error.

A comparison experiment should be run with supermod turned on, but without changing the values in R, so the effect of just the averaging can be seen.

Thank you for your insightful comments. Your points have motivated some major changes to the manuscript, and we have re-run most of the experiments.

The choice of α in the original manuscript was, as explained above, based on a set of experiments where we assimilated each satellite product with different choices for α. These experiments were validated against SLSTR Sentinel-3A SSTs. We found that for PMW SSTs assimilated through the supermod operator, the experiment with α = 0.3 validated better than the others. For the IR SSTs, α = 2 was the best option. We also expected that applying the supermod operator would allow for a lower observation error as the contribution from the representation error would be lower in this case.

With regard to the predictable relation to the error, we agree that there is a predictable reduction to the error of the background projected in the observation space with increasing footprints (by a factor 1/N where N=(1+2L)^2, given no spatial correlations in the background error). It is, however, less obvious to us that this reduction will be the same as the reduction of the error of representativeness.

To make the experiments easier to interpret we have rerun the experiments such that all satellite products use the same value for α, as suggested, and modified the manuscript accordingly.

P5/Table 1: for NPP and S-3A, you should write "validation only" or "not assimilated" for their alpha value. Don't just put a "-". I had to dig around in the main text to find out why alpha values were not given for these instruments.

All experiments use the same value for α in the updated manuscript. We have thus removed the column in question from Tab. 1.

P6/Figure 2: What does the change in colour mean?

The colors indicated temperature and should indeed have been accompanied by a colorbar and an explanation. As the purpose of the figure is to show the location of the drifting buoys (and the subdomain used for the spectral analysis), we have updated the figure to show the buoy locations in a single color.

P7/L150: There are other bias correction schemes, including online variational methods. These should be acknowledged here. https://onlinelibrary.wiley.com/doi/abs/10.1002/qj.3590 describes one such scheme and also contains references to other methods. I recommend having a look through this literature.

Thank you very much for pointing us to this literature. We have added some references to the manuscript in Sect. 2.3.

P8/Figure 4: I think this should be "bias" not "mean bias". "Mean bias" implies that you averaged several bias fields, which is not what the caption implies.

Indeed. We have changed the caption accordingly.

P8/L162: *"is less than 3 m/s during local summer daytime (defined here as May–August at 10–14 UTC). This wind speed threshold is less strict than that recommended in Donlon et al. (2002) but was chosen to ensure sufficient spatial coverage in the daily fields."*
I don't like the argument for halving the Donlon recommendation. Just wanting more data is not a good enough reason for keeping potentially bad data. Either switch to using a 6 m/s cutoff, or demonstrate quantitatively that using 3 m/s is scientifically acceptable.

The decision to lower the wind speed threshold in the bias correction scheme is based on the fact that diurnal warming events are relatively unusual in great parts of the region we are modeling (compared to the frequent occurrence at latitudes further south). Using a wind speed threshold value of 6 m/s will thus result in the elimination of useful observations from areas without diurnal warming, which again could result in degraded bias estimates, especially during cloudy conditions.

Fig. 3 shows an example of a situation where the removal of observations results in a degraded bias field. Here, biases are calculated for a day during a cloudy period using a threshold value of 3 m/s (Fig. 3a) and 6 m/s (Fig. 3b). We observe that the bias field produced with a threshold value of 6 m/s is even more patchy and contains an unrealistic warm bias "blob" (see the region enclosed by the black circle in Fig. 3b). That the bias field ends up like this is likely due to the high number of observations removed when we increase the threshold to 6 m/s (as seen in Fig. 4a and b). We found no indications that supported that diurnal warming events caused this region to differ between the bias fields produced for the different thresholds.

With a threshold of 3 m/s, we can retain many of the observations from areas without diurnal warming. However, with this threshold, we find that we do accept observations affected by diurnal warming, especially when the warming occurs in the southernmost part of the domain.

We acknowledge the need to reassess the wind speed threshold value. A threshold of 3 m/s is indeed too low. However, the recommended threshold of 6 m/s results in degraded bias estimates, especially when we have cloudy conditions. We are grateful that you brought this to our attention, and we have modified the paragraph where we introduce this threshold such that we state that this threshold should be reassessed. Additionally, we suggest exploring a coarser analysis grid since this may reduce the risk of degraded bias estimates. Due to the computational costs, we have not re-run the experiments with a bias correction scheme that uses a higher

threshold for the wind speed. However, our conclusions about the benefits of the scheme would not change.

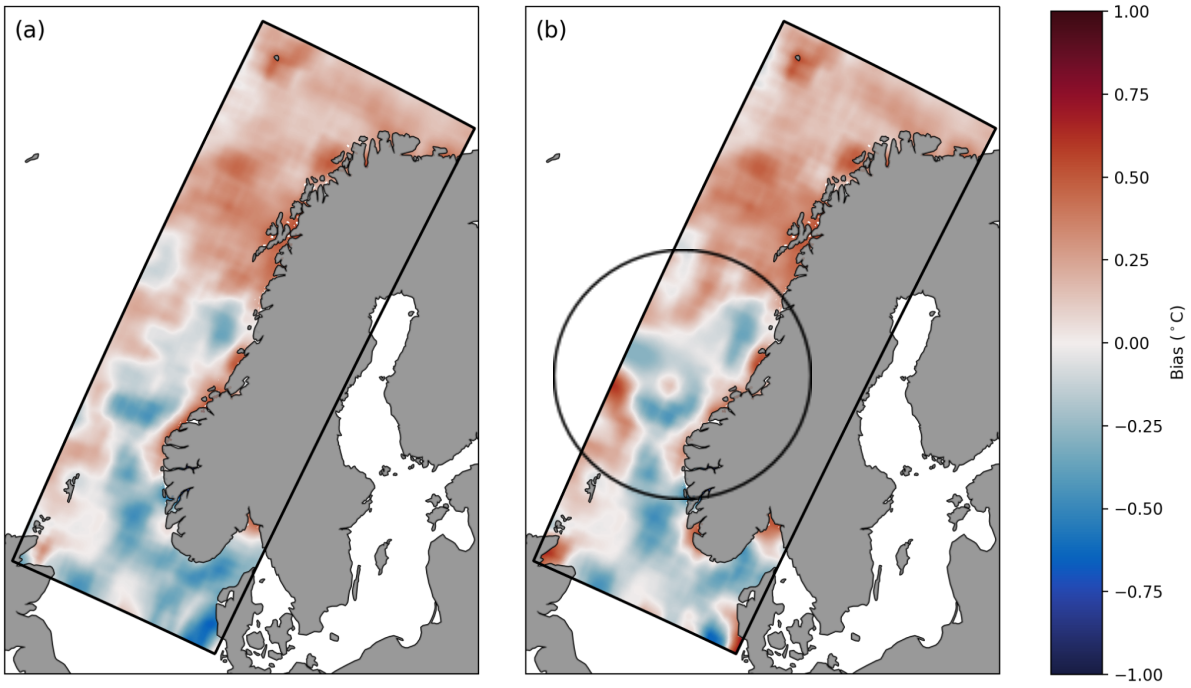

Figure 3: Bias calculated for 10 June 2018 with the threshold value for diurnal warming events set to (a) 3 m/s and (b) 6 m/s. Biases are calculated for AVHRR Metop-B SST.

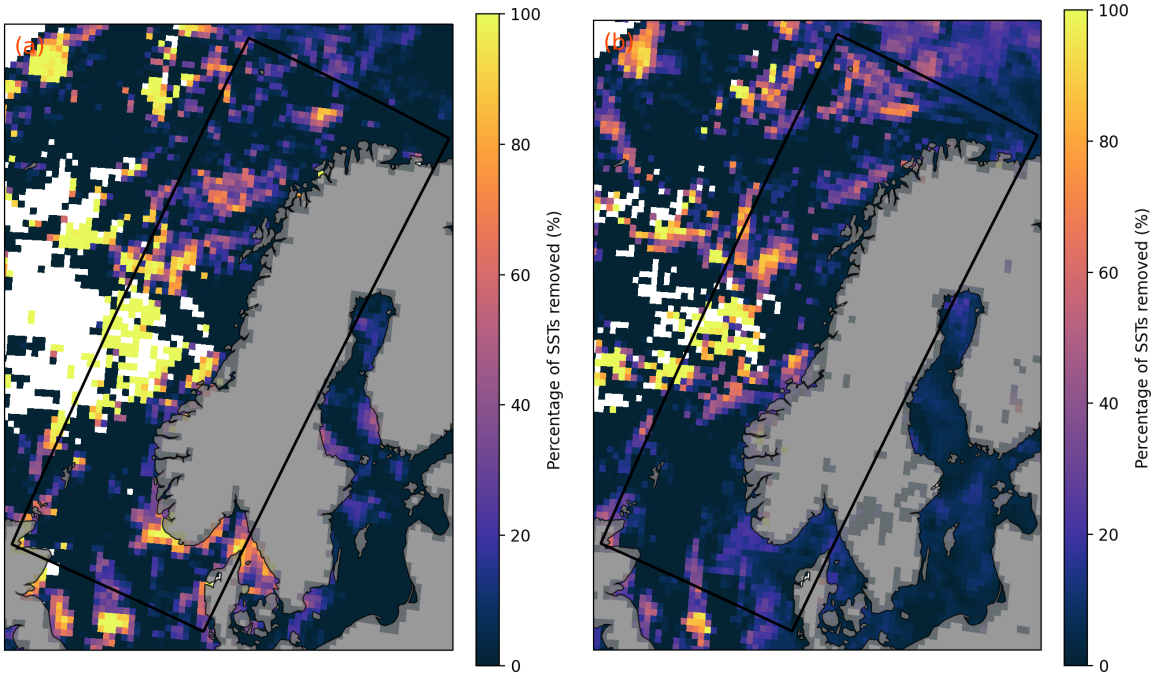

Figure 4: Percentage reduction in the number of observations used to calculate the bias for 10 June 2018 when going from a threshold value of 3 m/s to a threshold value of 6 m/s. (a) shows the reduction for the reference SSTs (SLSTR Sentinel-3A), while (b) shows the reduction for AVHRR Metop-B SSTs.

P8/L170: *“persistent for approximately 11 days”*
I don't agree with this statement. A large short lived bias will look the same as a small longer lived bias - there will be no way to distinguish them. In fact, oscillating biases with frequencies that are harmonics of 1/11 cycles/day will be invisible to this method. I don't think this invalidates the method. But the authors should be careful in describing what it represents.

Thank you for pointing this out. We have removed this part from the manuscript.

P12/Table 2: minor point, but 2.4e7 + 1.7e6 should give 2.6e7 when rounded.

Indeed. These numbers are, however, calculated and rounded for the observations within each experiment. The sum of the rounded numbers is different from the rounded sum of the actual numbers. As the value given in Tab. 2 is closer to the actual number of observations in this experiment than the sum of the rounded numbers, we have left it as it was.

P13/Table 3: *“Observations in coastal regions, where land emissions contaminate PMW SSTs, are excluded from the reference data sets.”*
I don't think these should have been omitted - it is still part of your domain, and the validation data is fine near the coast. You should show the stats for both the whole domain and excluding the coast.

We omit to validate coastal regions because we want to compare the different experiments such that we can say something about the impact of the assimilated observations. Experiments PMW1 and PMW2 do not adjust the areas close to the coast since there are no observations in these regions, and we observe that these experiments develop a strong bias in the coastal regions in the southern part of the domain (for PMW1, see Fig. 5a-c). If included, the reference observations from this region would dominate the error statistics since this strong bias does not occur in experiments assimilating observations here (for IR2, see Fig. 5e-g). To highlight the potential contribution of PMW SSTs to improved estimates of SST in the ocean model in experiments COMB1 and COMB2, we have chosen to focus on the regions where PMW SSTs are available. These regions are where an impact from including the PMW SSTs can be expected.

Figure 5d and h show examples of the difference between modeled and reference SST when we discard the coastal reference observations in regions where PMW SSTs do not exist.

We have added an explanation as to why we omit to validate the regions close to the coast at the very beginning of Sect. 4.2.

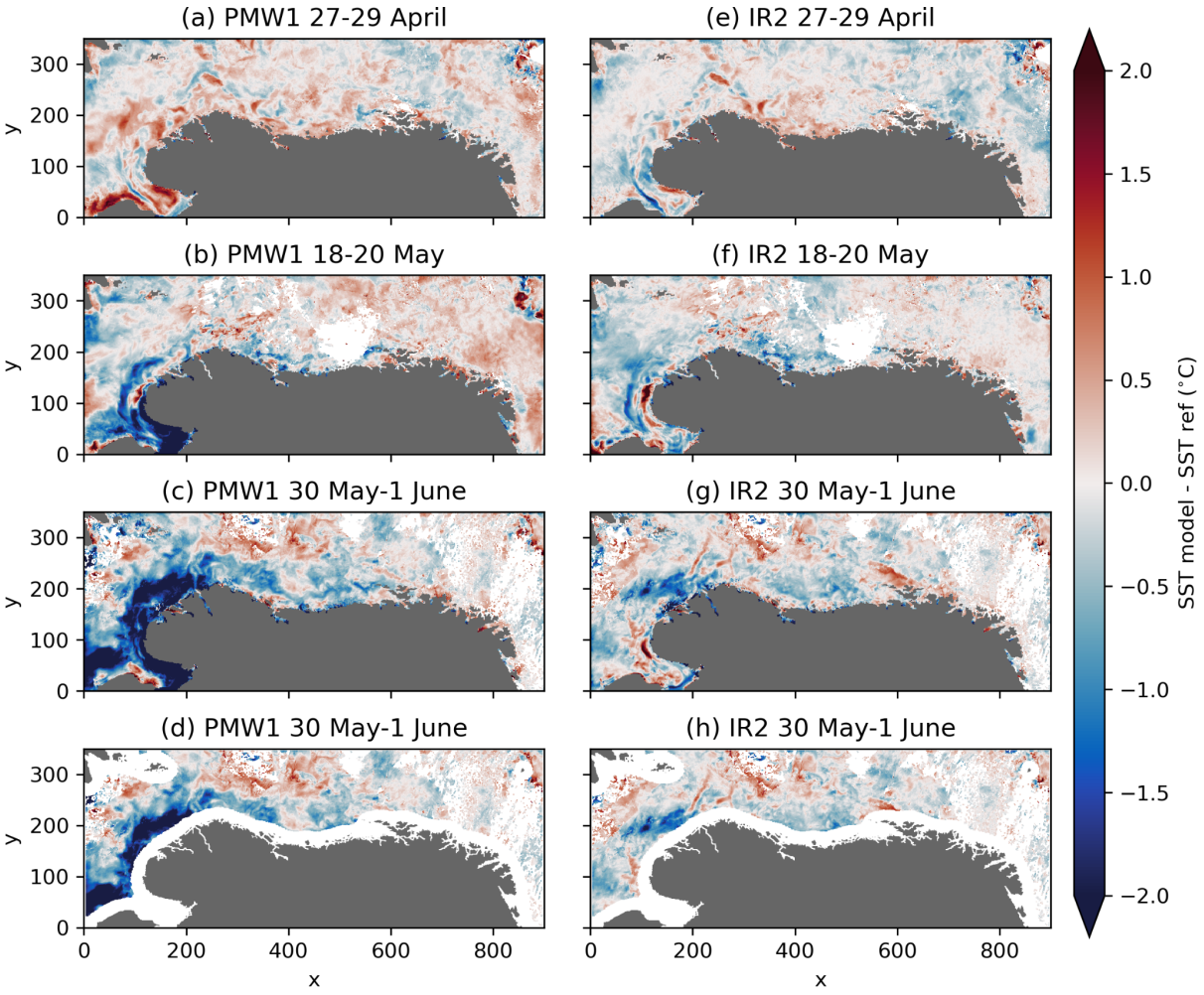

Figure 5: Difference between modeled SST and reference SST from satellites shown for (a-c) PMW1 and (e-g) IR2 for different assimilation cycles. (d) and (h) show the same fields as (c) and (g), respectively, but with discarded coastal reference observations in regions where PMW SSTs do not exist.

P14/Figure 7: Why are you only showing a "selected region" and not the whole domain? Also, where in the domain is this region? You need to add a plot indicating the location.

The figure was added to illustrate how the supermod operator works: increments become weaker and more smooth in PMW2. The figure also illustrates that it is not sufficient to only thin the PMW data and that you have to use the operator to spread the increment over the footprint. The reason why we chose a smaller region is that it made it easier to compare increment patterns in the different experiments. However, we agree that we should show the whole domain. The figure has been updated such that the whole domain is shown now.

P14/L264: *“PMW1, only with reduced increment amplitudes, indicates that the information provided by the additional SSTs in PMW1 is redundant.”*

They are similar, but they also have very significant differences. Thus this statement is far too strong. Notable differences are:
1) Plot (a) looks significantly less smooth; i.e., I suspect it has more high frequency information - as you would expect.
2) Negative increment top left of centre in plot (a) is absent in plot (b).
3) Negative zone in the bottom right is a notably different shape.

We have changed this statement in the updated manuscript. Now we discuss, in more detail, the similarities and differences. Additionally, the PMW2 increments with the higher α are much weaker and the statement about redundancy is certainly less justified now.

P15/L275: *"statistically significant at the 99 % level."*
What test did you use?

We have revised the method for showing statistical significance in the updated manuscript to ensure that the data sets that go into the tests are independent. A Wilcoxon signed-rank test is subsequently used to test for statistical significance. This information is now provided in the manuscript.

P15/L278: *"The better validation of COMB1 can also reflect that a smoothed version of a model field that resolves small-scale features generally validates better than the model field resolving these small-scale features. Small features captured by the model tend to have incorrect positions compared to the reference observations used for validation (Dagestad and Röhrs, 2019; Jacobs et al., 2021)."*
I find it very difficult to agree with the argument made here, which is a major flaw with the manuscript. Firstly the model is the same in all experiments, so worse statistics are just worse. I don't think the "double penalty" argument used here can be applied. And if it is, then it needs to be much better justified.

This statement is removed from the updated manuscript. With the original dataset, we did see more similar statistics when applying an averaging filter to the model SST fields before performing the validation. However, this is not the case for the updated experiments.

Also, a visual inspection of figure 7 seems to indicate that the unthinned internments contain power at higher wavenumbers than the thinned (but supermoded) increments. I think a spectral analysis of the increments - which I strongly suggest the authors perform - would confirm this. I therefore don't understand how apparently smoother data is generating more power at high wavenumbers in the model. I suspect over fitting, which would explain why the stats are worse.

Firstly, to confirm any overfitting in PMW2 and COMB2, we used the diagnosis presented in Desroziers et al. (2005) using the innovations from these experiments. This confirmed that the α used for the PMW SSTs in these experiments is indeed too low. The updated manuscript now uses α = 2 for all experiments. Potential overfitting is also evaluated using the same method for all of the new experiments, and the results indicate that α = 2 is too high for all of them. This choice for α leads to a deterioration of the error statistics of PMW2 and COMB2. The power spectra, on the other hand, still shows that PMW1/COMB1 has less variability than PMW2/COMB2 at spatial scales smaller than ~120 km. This may seem contradictory, given the fact that the increments of PMW2 indeed are smoother than those of PMW1. However, the detailed increments often reflect the removal or dampening of small-scale features present in the model background SST fields (See Fig. 1 and Fig. 2 above). We have added a comment on this to the discussion of Fig. 7 in Sect. 4.2 to ease the interpretation of the results.

P15/L285: *"We find that the cloudy period has an RMSE of 0.443 ◦ C and 0.428 ◦ C in IR2 and COMB2, respectively. The bias is −0.110 ◦ C in IR2, while it is slightly reduced to −0.098 ◦ C in COMB2."*
What is used for valiation here. Your satellite validation data will also be affected by clouds and this needs to be accounted for. Also - these are not big changes, are they actually significant?

Thank you for pointing this out. Reviewer #1 also pointed out this issue, and we have implemented a new method for assessing the performance in cloudy and clear-sky conditions. The error statistics of the background for each cycle are now calculated by using reference observations where we had clouds during the previous cycle (which is the cycle that created the initial conditions which the background was initiated from). Similarly, clear-sky conditions are evaluated by evaluating areas where we had good IR SST coverage during the previous cycle. The results are also tested for statistical significance in the same way as for the other error statistics.

P15/L301: What do you mean by "this" - just the data you use here, or all PMW data.

We refer to the PMW dataset we use in this study. However, the paper we reference in this regard, Alerskans et al. (2020), states that radio frequency interference (RFI) is a problem for both SST and wind speed retrievals for certain frequencies within the PMW frequency range. The problem of RFI is thus not necessarily restricted to the AMSR-2 data we use. Furthermore, data within many of the areas where we would expect such interference is already flagged as having inferior quality in the data set we use. Thus, whether or not it is a problem in other PMW products probably depends on the algorithms used to detect such interference for each product in question.

Since we re-ran all experiments using PMW SSTs, we also chose to apply a simple filter removing PMW SSTs within 75 km of oil rigs (from a predefined list of positions). This part of the manuscript where we mention the problem of RFI is thus moved to Sect. 2.2.

P16/L307: "wavenumber" not "frequency" as you are dealing with space not time.

Thank you for noticing this, it has been updated.

P16/L308: I'm not sure you should talk about the gradients when the spectra are of the actual field.

That is a good point. We have changed this in the updated manuscript.

P17/L313: *"A disadvantage of detrending and windowing is that these methods contaminate the largest scales in the spectral analysis."*
Tapering (what you call windowing) lowers the resolution of the spectrum, but it doesn't "contaminate" the larger scales.

We have reformulated this statement.

P17/L317: Please use the more common "power" spectrum (or power spectral density) rather than variance spectrum.

We now use "power spectrum" instead of "variance spectrum".

P17/L319: "Ricard et al. (2013)"
A reference isn't good enough here. I'd like to see an actual description of how the binning was done.

The manuscript has been updated to include more information about this process.

P17/L330: I think you need to show the spectrum of the increments as well. The PMW1 increments in figure 7 certainly looked to have power at higher wavenumbers than PMW2.

We agree that there will be more power in the spectrum for the PMW1 increments. However, as explained previously, that the increments are detailed (and stronger) does not necessarily result in a more detailed SST field. Sometimes it does, but other times the strong and detailed increments represent a removal or smoothening of detailed structures in the background.

To illustrate this, we have plotted the increments, the background, and the analysis in PMW1 and PMW2, and the results are shown in Fig. 6 and Fig. 7, respectively. For PMW1, we see that many of the detailed increments represent removal of structures from the background state. Examples are:

- the red increment structure at indexes (x=350, y=320),
- the blue pattern at (370, 290),
- the blue pattern at (400, 330),
- the blue "blob" at (520, 290),
- the dark red increment at (540, 220),
- and the red pattern in the upper, right corner.

The increments in PMW2 are smoother and act more as an adjustment of the mean of a larger area.

Since our goal is to figure out if the SST in the background state becomes smoother, we calculate the spectrum using the SST from the background states.

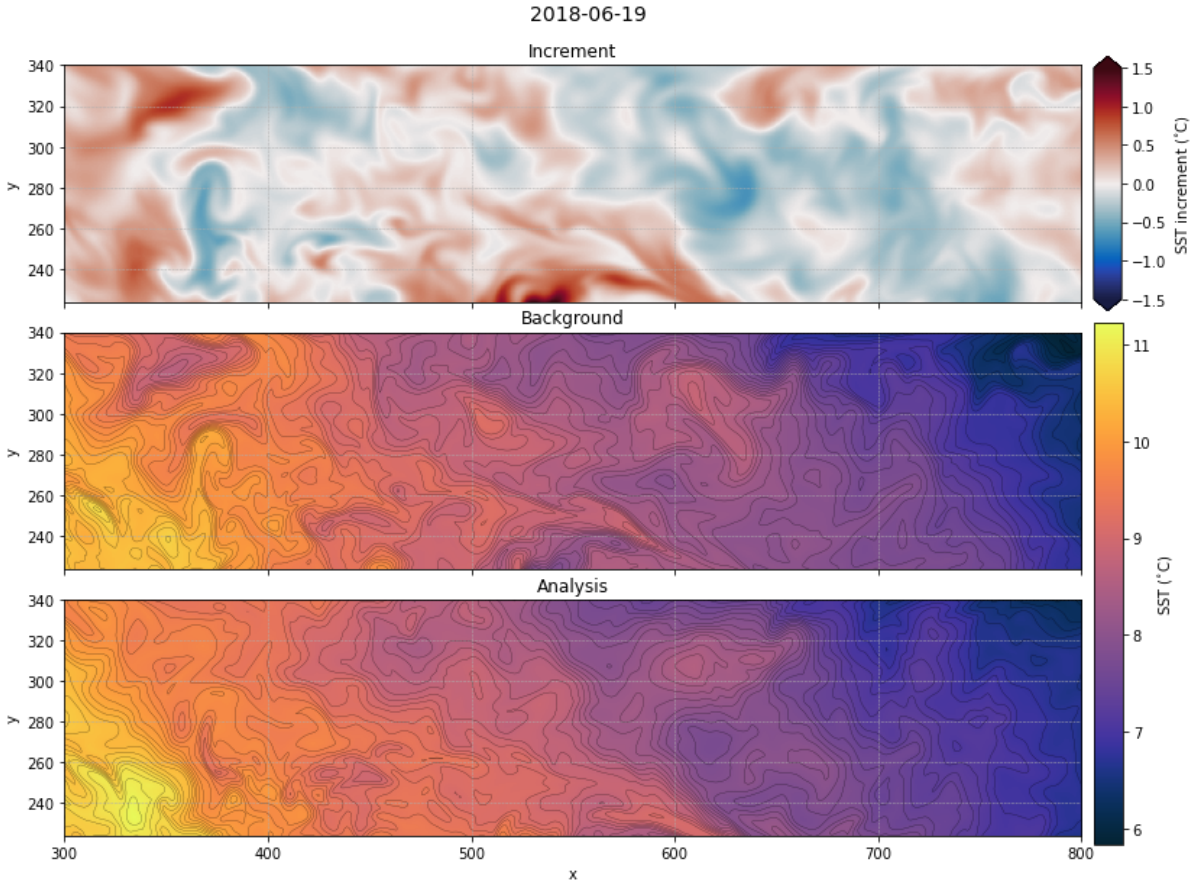

Figure 6: PMW1 increment (upper), background (middle), and analysis (lower). Average over all time steps 19 June 2018. For the background and the analysis, black contours are drawn every 0.1 C.

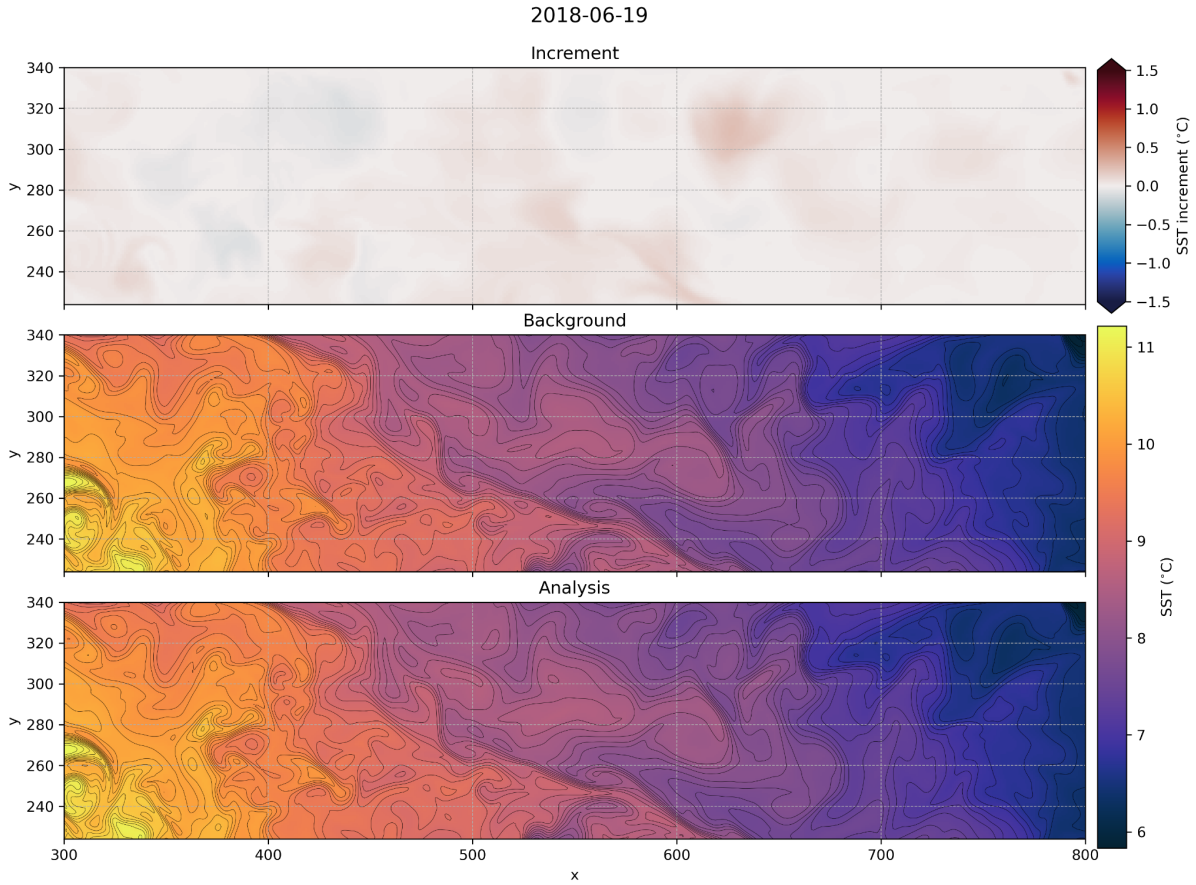

Figure 7: PMW2 increment (upper), background (middle), and analysis (lower). Average over all time steps 19 June 2018. For the background and the analysis, black contours are drawn every 0.1 C.

P18/Figure 9: I think what your are showing here is the power spectral density (which can be estimated as the square of the DFT of the data). In which case the units on the y axis are wrong. It's a density and should have units of Deg C per wavenumber (or "Deg C km", in your case).

Following Denis et al. (2002), the two-dimensional spectral variance array has dimensions of deg C squared. The wavenumber associated with each element in this array is adimensional, and can be written as

$$\kappa = \sqrt{\frac{m^2}{N_i^2} + \frac{n^2}{N_j^2}}$$

where m and n are the grid cell indexes in both horizontal directions, and Ni and Nj are the total number of grid cells in each horizontal direction. When we perform the binning, we thus end up with a one dimensional spectrum with units of deg C squared. So the spectrum is in “grid cell- or pixel-space”.

When we present the results, we refer to the wavelength in km instead of the adimensional wavenumber to ease the interpretation. We have converted the wavenumber to a wavelength through

$$\lambda = \frac{2\Delta}{\kappa}$$

which is also found in Denis et al. (2002). We have added some additional information about this in the manuscript.

P19/L350: "RMSE and bias of COMB1 do not reflect that this experiment is a better quality product."
I am unconvinced. The two systems are at the same resolution and using the same verification data, so I would expect to see improved statistics not degraded ones. I think you are overfitting the data because you have reduced the observation error too much. You need to convince me otherwise, by eliminating this as a possibility.

This statement is changed in the updated version.

References

Alerskans, E., Høyer, J. L., Gentemann, C. L., Pedersen, L. T., Nielsen-Englyst, P., and Donlon, C.: Construction of a climate data record of sea surface temperature from passive microwave measurements, Remote Sens. Environ., 236, 111 485, https://doi.org/10.1016/j.rse.2019.111485, 2020.

Denis, B., Côté, J., and Laprise, R.: Spectral Decomposition of Two-Dimensional Atmospheric Fields on Limited-Area Domains Using the Discrete Cosine Transform (DCT), Mon. Weather Rev., 130, 1812–1829, https://doi.org/10.1175/1520-0493(2002)130<1812:SDOTDA>2.0.CO;2, 2002.

Desroziers, G., Berre, L., Chapnik, B., and Poli, P.: Diagnosis of observation, background and analysis-error statistics in observation space, Q. J. Roy. Meteor. Soc., 131, 3385–3396, https://doi.org/10.1256/qj.05.108, 2005.