Authors' response to Referee Comment 1

We would like to thank the referee for taking the time to carefully read our manuscript and provide a large number of suggestions and comments which we find very useful for the present revision of our manuscript.

Please find our detailed responses to all specific comments below, and note that while we have followed the referee's advise on most of the items, there are a few upon which we have not acted. Initial page and line numbers below (in bold) and comments (in italics) are repeated from the referee's document. This is followed by our response (in regular font) and, when relevant, reference to where changes can be found in the revised submission (in italic bold).

P3L29 and elsewhere

often "grid(s)" is used when "grid cell(s)" is meant. Also, some- times "nodes" is used instead. I recommend to use "grid cell(s)" consistently (where that is meant, of course). This also holds for the Supplement.

Where applicable we have replaced "node(s)" and "grid(s)" with "grid cell(s)" (also in the Supplementary Information document). An example where "grid" was not modified is when referring to a "stereographic grid".

P4Eq7

I suggest to make it explicit that d_o and d_m are not single scalars but sets, if I am not mistaken, by writing the right-hand-side as "$\max(\max(d_o), \max(d_m))$".

Even though the contents of Eq. 7 is not affected, we have elected to follow the referee's suggestion and modified the equation as recommended. ***P5L6***

P5Eq8

It seems that statements like "$a^+ = 0$ elsewhere" and "$a^- = 0$ elsewhere" are missing in the upper and lower equation, respectively.

The referee is correct, and Eq. 8 has been rewritten accordingly. ***P5L14***

P7Eq17

I am somewhat irritated by this equation. For example, when I substitute i_k^n (bottom left) into the upper equation, the first term in the brackets becomes $1 + k \cdot (n + 1)$, which doesnt seem to make much sense. Isnt i_k^n supposed to stay the same when the sums are evaluated, that is, should the indices be different?

The referee's irritation regarding Eq. 17 is highly justified. We have taken two actions related to this issue. First, in our original manuscript I as defined by Eq. 16 is a grid cell quantity. Since all other quantities for the grid cell level are in lower case, this was unfortunate. Hence, we have replaced I by λ in the present revision. Second, the reviewer rightly rejects the use of *e.g.* i_k^n in Eq. 17, the correct here is i^n. The equation has been corrected accordingly. ***P8L3***

P8L3-9

It might be OK not to repeat the algorithm for the FSS displacement, but at least a qualitative description of how that quantity is derived from the FSS should be provided.

We have rewritten Sect. 2.3 to provide more general information on the FSS metric in the first paragraphs, and we also provide an approximate expression for the relation between FSS values and FSS displacement lengths in the final paragraph. ***P7L16-23, P9L16-17***

P9L16-17

"the resulting displacement metrics are also reduced substantially from the Reference case to the Modified case, due to the added ice areas proximity to land."; Is this sentence really saying what its supposed to say? After all, they are still increasing, only much less.

The referee is correct, and the sentence in question have been rewritten accordingly. ***P10L21-22***

P10L4-7

It seems worth mentioning that the Hausdorff-type metrics do not require remapping, although it seems OK to do it in this study to ensure consistency. This could also be mentioned in the discussion part

The referee is correct that Hausdorff-type metrics do not require remapping. The main contrasts between our approach and that of some other investigations is that we treat the ice edge as being composed of grid cells, rather than one-dimensional curves. We have added a paragraph on this topic (the second paragraph in Sect. 2). Moreover, while it is possible to define displacement metrics also for sets of grid cells given on different resolutions and projections, there are then complications related to representativeness that we find to be somewhat beyond the scope of the present study. ***P3L27-29***

P10L23-32

Here I was surprised that the relation between $\widehat{D^{IE}}$ and D^{IE} is not mentioned, and also not the relation between $\widehat{D^{IIEE}}$ and D^{IIEE}. Likewise, its worth to highlight already that D^{IIEE} and $\widehat{D^{IE}}$ are very similar. You elaborate on this only in the next section, and I think this is an interesting outcome that gives confidence about the robustness of these two metrics which are technically derived in quite different ways.

As suggested here by the referee we have added a section (second to last in Sect. 4) where results for various metrics from the two forecasts are mentioned. ***P12L8-13***

P11L22-26

What can be concluded from the comparison of the two different observational products? Can this help to understand the relatively large errors that are present already in the initial states? It would be good to comment on this.

The referee is correct about the impact of the contrasts between the assimilated microwave data and the ice chart data used for validation. We have added a sentence about this in the paragraph in question, and also in the following section. However, we refer to initial differences as 'deviations' rather than 'errors' since the two observational products in question have their separate strengths and weaknesses, so the 'truth' is not known. Finally, additional results from the comparison between the two observational products are now given in the Supplementary Information. ***P13L9-11,18-23, Sect. S2, Tables S1,S2, Fig. S3***

P12L8

"This was to be expected"; Actually, I would not have expected such a close match, given the considerably different approach to derive these two metrics.

We admit that the expected relationship between displacement metrics should have been explained more carefully. In the present revision we have included a discussion of idealized cases in the beginning of Sect. 6 which should shed light on this topic. ***P13L28-33***

P12L18

"50 such pairs" -> "50 out of 105 pairs" (correct?)

Yes, it's 50 out of 105 pairs. This is stated explicitly in the revised manuscript. ***P14L18***

P14L5-6

Regarding the maps, these would be examples of past performance rather some kinds of averages, which I wouldnt know how that should work, right? Or maps showing the errors for the latest previous forecasts (making use of the slow decorrelation)?

Our recommendation is due to the latter, *i.e.* the long decorrelation time scale. In order to explain this better, we have rearranged Sect. 6.3 and rewritten the sentence in question. ***P16L20-22***

P14L14-18

I have difficulties to understand this paragraph on the usefulness of providing FSS in addition. I suggest to either explain a bit more, or to remove this paragraph.

The sentence concerning steepness of 0.5-crossing was not documented, and may thus be incorrect.

This sentence has been removed. However, the application of FSS for examination of systems with different resolutions is at the core of this metric, and has been described thoroughly in papers that we cite, see *e.g.* Roberts and Lean. This is also stated in the Introduction section of the present manuscript. Based on suggestions from another referee the presentation of the FSS metric has been re-arranged in this revision.

P15L1-3

Is the Palerme et al. paper published now? Its not ideal to base an important final recommendation partly on a not-yet-published paper.

Palerme et al. is not yet published, but a revised manuscript based on a 'minor revision' recommendation has been submitted. However, we disagree that our recommendation is partly based on this study. Palerme et al. was mentioned here for context. Nevertheless, we have moved this sentence to the Introduction section, where it fits nicely in a paragraph where relevant literature is listed. ***P2L20***

P15L3-4

"We have shown that the deterioration in the forecast quality is moderate for these lead times"; Again, I think there should be some discussion on why there is such a relatively large initial error (which is partly responsible for this slow initial error growth, I would say).

A discussion on the impact of initial errors, or rather deviations, is provided in Sect. 5 in the revised manuscript, see our reply to item **P11L22-26** above.

Figure2

Is A^- and A^+ the wrong way around here? Shouldnt A^+ be the part where the model/forecast has too much ice?

The referee asks if there is an error in the color shading in Fig. 2 in our original submission, and we have indeed made the mistake that the referee has spotted. We are very grateful that the referee pointed us to our mistake. In the revised manuscript the error has been corrected. We can add that we double-checked Fig. 5 (Fig. 3 in the original submission), and found that this did *not* contain the same mistake. ***Fig. 4***

Figure5

A statement on the units of the y-axis is missing.

The units of the y-axis is now given in the caption. ***Fig. 7***

EqS2-S4

It appears strange to me to use the areas (a^{ia}) as weights when averaging over the different segments the edge consists of. Wouldnt it make more sense to use the lengths l as weights? In case of S3, and neglecting A_0, this would yield simply $D... = \sum a / \sum l$. Also, for the same reason, the term A_0 seems a bit arbitrary: this one would converge to zero for increasing resolution, right? I am also suspecting that this awkward weighting is the reason why the hat-versions of D^{IIEE} are by such a large factor larger than those without hat.

The application of area weights was introduced in order to highlight effects of the geometry of IIEE areas, as stated in Sect. 2.2.2. With the referee's suggestion (*e.g.* $\sum a^{ia} / \sum l^{ia}$) the metrics would essentially give the same information as the original $D^{^{IIEE}}$ metrics: consider the fraction $\widehat{D^{^{IIEE}}_{_{AVG}}} / D^{^{IIEE}}_{_{AVG}}$ in the three idealized cases we present. For $\nu = 1/4$ the resulting fractions are $1.5, 1.7$ and 1.35, respectively. Adopting the referee's suggestion we find the set of corresponding values to be 1.38, 1.36 and 1.36. For $\nu = 4$ the resulting fractions are $3, 2.5$ and 2.3, while the referee's alternative yields fractions of 1.17, 1.13, 1.13. Moreover, the term A_0 is not arbitrary: in the case of two identical ice edges, dropping A_0 will lead to an ill-defined value for $\widehat{D^{^{IIEE}}_{_{AVG}}}$ since with no A_0 it becomes $0/0$. Note that A_0 is the area of all grid cells where the products overlap, the sentence in question has been rewritten to make this clear. ***S-P1L17-18***

Technical corrections

All of these items have been corrected according to the referee's advice.