Thank you to the Referees for taking the time to read through this paper and add your comments. I have some general comments and a short reply for each point made in the comments, starting with the short comments, then replies to the anonymous referees.

General Comments:
Judging from the general comments of both referees, I suspect that the title is misleading. The goal of this paper was never to present a new method, but to demonstrate using a simple and transparent method that the POLCOMS-ERSEM performed admirably in a region that is historically important for international marine policy. The simplicity of the point to point matching and the linear regression is part of why we were impressed with the results of this model.

An issue with Marine Ecosystem modelling, as opposed to circulation models, is that even the state of the art models have a relatively large margin of uncertainty and can struggle to reproduce nature. Typically, Marine Ecosystem models reproduce trends loosely and we felt that this method should be commended for demonstrating the model's success in (more or less) reproducing in situ measurements over multiple decades.

We've also made many other changes to the text:

- Changed the title and abstract to focus on this being an example instead of a presentation of a new method.
- Adding references to other point-to-point matching papers.
- Emphasised the three dimensional nature of the in situ data and note that extra layer of difficulty of interpolating ocean data in three dimensions with a complex water column.
- Emphasising policy as motivation for this work, especially the importance of demonstrable model quality in the North Sea.
- Added a more detailed description of the model, especially the ERA 40 reanalysis and the river influxes.
- Added a better description of the difference between time cut and time granularity to the methods section.
- The tables 1 and 2 were split into five table, or one per dataset, containing all 10 permutations of granularity and time selection.
- Re-working the results section, especially the description of the time series plots, which led to some confusion.
- To reduce confusion, we replaced all mentions of the term "entire region data" with "Mixed layed depth (MLD) averaged data".
- Many small changes elsewhere.

Replies by Lee de Mora in GREEN

Reply to Short Comments:
H. Riede
hella.riede@mpic.de
Received and published: 5 September 2012

HR: For some other figures on the impact of sampling in time and space, http://www.geoscimodel-dev.net/3/717/2010/gmd-3-717-2010.html Fig. 6 and Fig. 8 may be of interest to you (capturing full model information and evading avoiding artefacts by point-to-point sampling).

Lee: This is indeed of interest to us. Figure 8 of this paper demonstrates the value of on-line over interpolation of sampling in atmospheric OH mixing. In the case of marine biogechemistry, the response time and rates of change are much slower and can usually be measured as a daily rate.
Secondly, I fear that this method may be technically impractical for this model. It takes a month of running time on our Cluster to produce our 45 year hindcast. If we were to extract data that was matched to the hour (in this model run) with many million in situ measurements, it could easily double the runtime required. Finally, logistically, this work was produced using a pre-prepared model run, and it was not in the scope of the work to perform additional model runs.
I have added the following paragraph to the end of the "point to point matching" subsection:
"Techniques similar to this point-to-point matching have been used elsewhere in geoscientific models, but are rare in marine ecosystem modelling. For instance, in (Jockel et al), a dataset was created during the model run that matched a specific flight-path with the highest model time resolution available. Unfortunately, the typical time required to produce a 45 year POLCOMS-ERSEM hindcast is order one month and implementing this technique could double the run time requirement. For this reason, runtime methods of data recording were beyond the scope of this work."

Reply to Anonymous Referee #1:
Anonymous Referee #1
Received and published: 24 September 2012

AR#1 General comments
The authors introduce a methodology for the comparison of model results with sparse in situ measurements. The suggested point-to-point method seems to be a canonical method to do this but is firstly not often used and secondly as far as I can see not been systematically investigated. The authors use this method for the comparison of ICES data to their North Sea Model ERSEM/POLCOMS. The region of the data set and the model area do not coincide. Thus it is strongly recommended to do some kind of snapping of data and model results. This is normally done by restricting the model results to be compared to the region where data are available. The authors extended this methodology by a point-to-point snapping in space and time. This leads in most cases of their example to better accordance than the comparison within the whole region. But this is not a proof that this methodology has a higher evidence. It is true for the here presented model. It is obvious that a mathematical proof of the methodology in a widely generalized way can not easiliy been done. But it could be verified for a combination of model and data where the quality of the model is well known. This can f.e. been done by taking model results and sparse data sets derived from these model results modified by a Gaussian blur and compare these. This could be done for several randomly chosen data sets and a test can be established if the suggested method leads statistically to a better coincidence. An example of such an investigation can be found at the end of Jolliff et al. 2009 where the used data set is a modified model result with controlled statistical parameters.

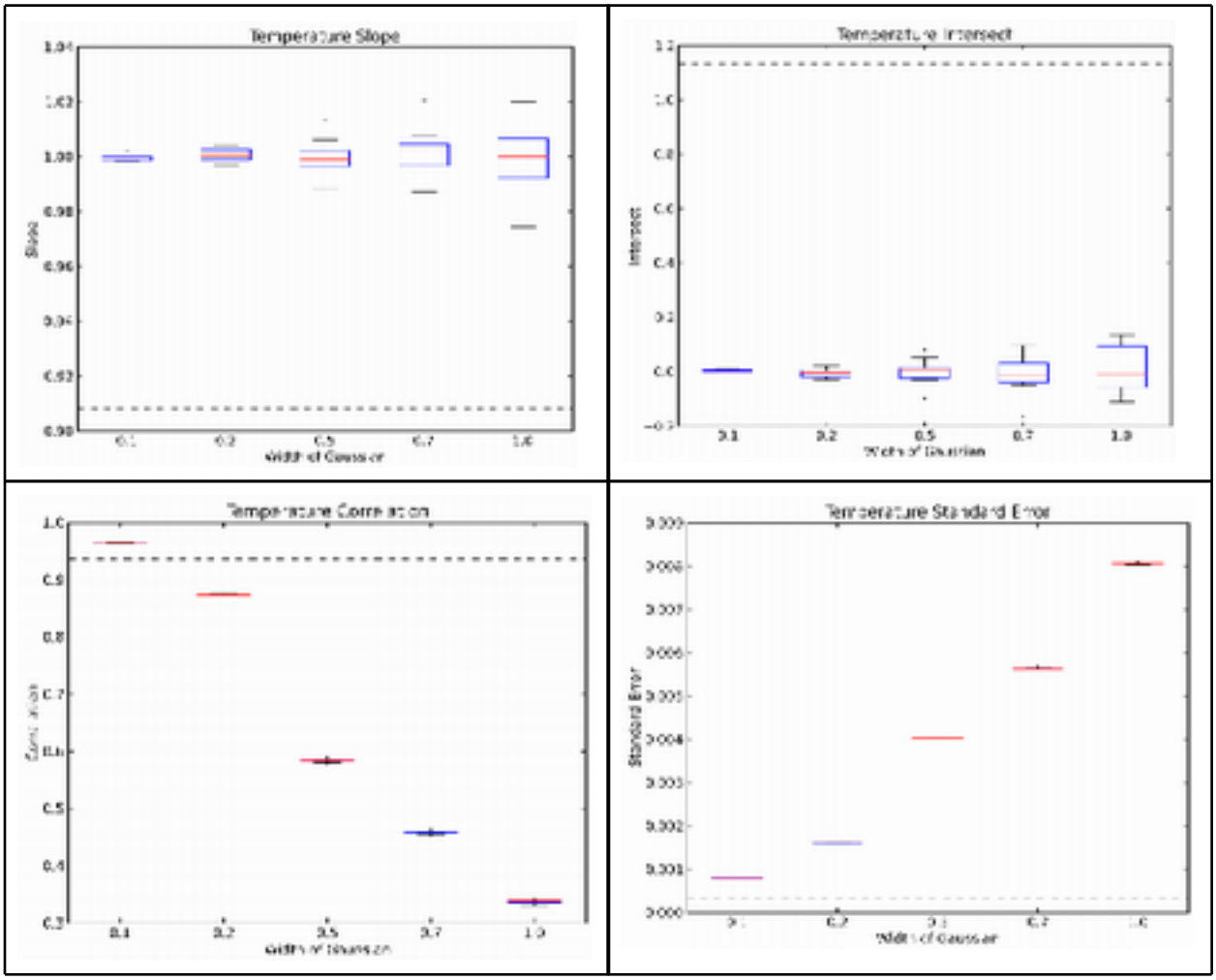
Lee: AR#1 suggests to take a selection of the model point to point datasets, modify these by blurring each point randomly with a gaussian distributions, then comparing those. However, if we were to blur the data as in Jolliff 2009, it would have the same effect as increasing the size of each model pixel, further smoothing out an already smooth dataset. Furthermore, any shift observed in linear regressions would clearly be the change that was implemented. A gaussian blur shifts the model data to make it less like in situ data. If we wanted to make the model data look more like in situ data, random noise could be added. We performed the following analysis as a test:

Each of the five model point-to-point datasets (Salinity, Temperature, Chl, n3, p4) in the North Sea were subsampled, and a gaussian noise was applied to the model data, ie

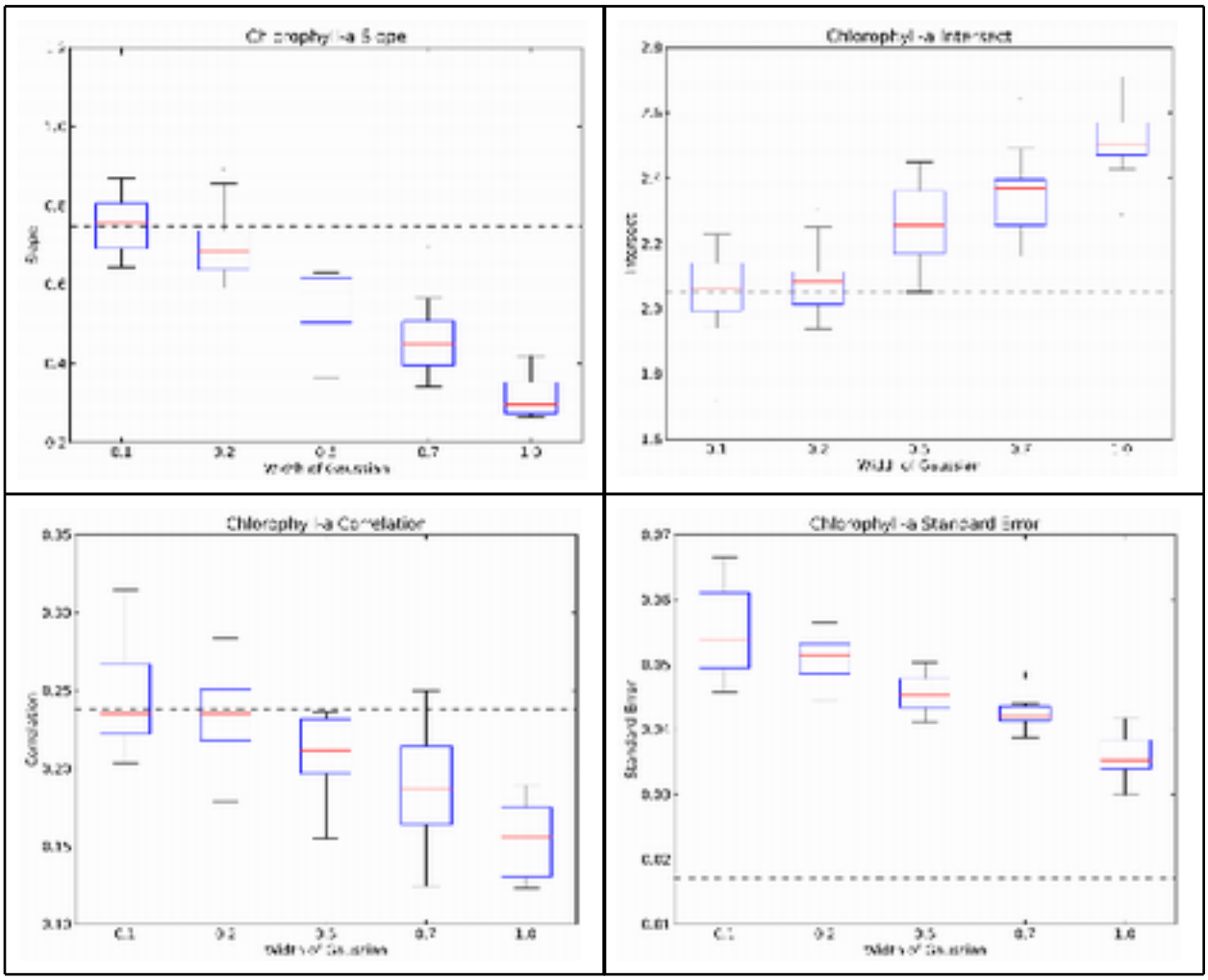
- noisey point = original point + original point x Gaussian (mean = 0, width)

where we varied the Gaussian width between 0.1 and 1.0. Then a linear regression was applied times to 10 independent model subsamples against the noisy model subsamples. The linear regressions output parameters (slope, intersect, correlation, p-value, standard error) were plotted in "Box and Whisker" diagrams.

These figures (below) show exactly what you would expect from comparing a dataset to a noisy version of itself. As the noisiness (width of gaussian) increases, the two datasets diverge. For this reason, we did not find it useful to compare a dataset to a slightly altered version of itself.

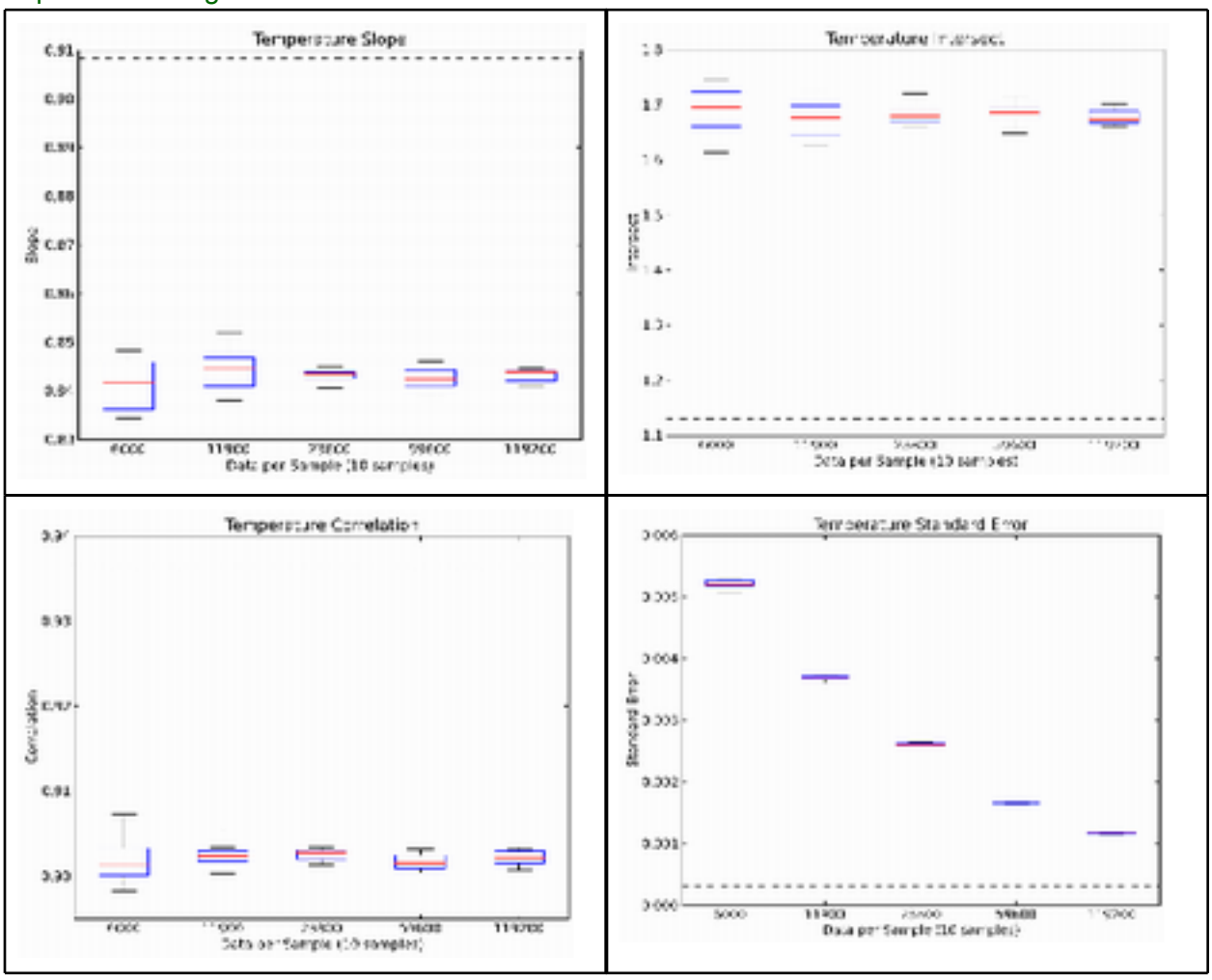


A second exercise was performed, where an increasingly noisy subsample of the model was repeatedly compared against the the matched in situ sub-sample, with a range of data per sample. The following plots were created, for Chlorophyll (below). The horizontal line is the matched model against in situ with no noise. By adding noise, this exercise tilts the model such that as noise increases, the linear regression tends to a line of slope 0 and low correlation. As chlorophyll was the dataset where we had previously seen the worst correlations and fits, we were not expecting an increase in noise to have a large impact on the linear regression. However, as these plots demonstrate, increasing in noise does in fact make the model look less and less like the in situ data, even in the case where the model performed the poorest. Interestingly, the Standard error decreases with the width of the Gaussian.



Finally, a third exercise was performed. Here, a subsample of the noisy model was repeatedly compared against the in situ, with a varying number of data per sample. Here, a constant gaussian width of 0.1 was used to create the following temperature plots. The subsampling was performed first by randomly selecting 10 independent samples of the total dataset. The selection of 10 independent samples was performed 5 times, and the size of the subsamples were varied between: 1/200th, 1/100th, 1/50th, 1/20th and 1/10th of the total dataset (rounded to the nearest 100 data points). i.e. if the total dataset had 200 000 data points, then the follow subsamples were prepared: 10 x 1000, 10 x 2000, 10x4000, 10 x 10000, 10 x 20000. The subsamples contained independent measurements to the other samples of the same size, but may overlap with subsamples of a different size.
These figures (below) show the slope, intersect, correlation and standard error for in situ temperature against noisey model temperature. As in the paper, we demonstrated the model performs very well at reproducing the in situ temperature measurements. The dashed horizontal line is the linear regression of the unblurred full sample. It should be noted that the slope, the intersect and the correlation all worsen when comparing a blurred sample to an unblurred

sample. Furthermore, there is no indicated that larger datasets would result in better agreement. This is expected as the quality of the model data is independent of the quantity of in situ measurements. The standard error is also worsened by the noisyness, but this is more likely due to smaller sample size than the effect of the Gaussian blurring. We conclude that the point to point matching leads to a better coincidence.



AR#1: This might not be the ambition of the authors who might possibly show the results only for the presented model. In this case the title of the paper is misleading. It should be at least extended by the term "an example". If the paper shall focus on this special example, the text should be restructered. The subsections of section 5 should be combined so that one has a clear view on the different representations of the results (tables and figures for the considered variables).
In summary: the paper shall either focus on the example -in this case the title is not appropriate- or the paper should focus on the methodology, then a general -possibly empirical- investigation is needed.

Lee: As mentioned above, I agree and we've changed the title to "How should sparse marine in situ measurements be compared to continuous model data: An example" in order to focus on this paper as an example of this method for marine ecosystem modelling.
Section headers 5.1, 5.2, 5.3 have been removed, and as suggested by AR#2, we are in the process of reworking the entire Section 5.

AR#1 It should be mentioned that similar methods are used even if they are not published, fe: 'Point-to-point matching' is discussed in Genkova et al.(2004) Point-to-point comparison of Satellite and Ground-based cloud properties at the ARM Souther Great Plains Central Facility and Comparison of ASPEN Modeling System Results to Monitored Concentrations at http://www.epa.gov/ttn/atw/nata/mtom_pre.html 'Snap-to-Grid' is a standard method in the ArcGIS product family.
Lee: We are aware that these methods exist elsewhere. However, these references relate to atmospheric and terrestrial modelling. This methodology doesn't appear to be used extremely frequently in marine ecosystem modelling, possibly because of the dominance of satellite products used in large scale model validation and the scarcity of large multi-decadal in situ databases. Secondly, these methods tend to focus in two dimensional datasets. We've changed the introductory text to be very specific about the three dimensional marine biogeochemical modelling aspect of this work.

AR#1: Specific comments
AR#1: Sec 2 (2314): It is described that the POLCOMS area cover a far larger area than the ICES region from which the in situ data were taken. It is not described whether in the later section of the article the POLCOMS output already has been reduced to the ICES area when talking about the entire model. If this reduction was not done before investigating the 'snap-to-grid' matching, the conclusion of this article is already obvious: 'Model results and in situ measurements always match better if the domains of both are matching better'.
Lee: The North Sea region was chosen because of the quantity and regularity of data there, but also because of its international importance in policy driven science. Furthermore, this region is distance enough of the edge of the AMM region to avoid the influence of open ocean boundary effects. The text has been changed in many places to reflect the importance of policy in the motivation of this work.
Secondly, I agree that the conclusion of the paper does indeed seem obvious: model results and in situ measurements always match better is the domains are matching better. However, in GMD-5-223-2012, (doi:10.5194/gmd-5-223-2012) Saux-Picart *et. al.* demonstrated that this is not the case: The current generation of POLCOMS-ERSEM do not perform equally well on all scales. In fact, they found that the model performed better on larger spatial scales than smaller ones for Chlorophyll, with relatively poor skill when matching against Satellite data on small spatial scales (in their case, "small spatial scales" were order 24km sided squares). For this reason, we were surprised to find the pixel-by-pixel matching against in situ data agreed as well as it did using such a simple method.

AR#1: Sec 4.1 (2316) line 21: Building the mean of multiple in situ measurements falling into

the same grid cell leads also to a bias. This can easily be understood: Assuming the measurements for the entire region are given by the numbers from 1 to 10. The entire region may be subdivided into two grid cells. The first 3 measurements may snap into the first cell, the rest into the second. The mean of the total region becomes biased: mean(1,2,..,10)=5.5 but mean(mean(1,2,3),mean(4,5,...,10))=mean(2,7)=4.5.
Lee: It is quite rare that more than one in situ measurement appears in the same box on the same day. However, this does happen and taking the mean seem unavoidable, would it be better to use multiple copies of the same model pixel?

AR#1: Sec 4.1 (2317) line 5: It is not clear what the reason for the avaraging of the data set is? Averaging over the year or a season does the same with a data set as averaging over a geographical region. Thus, the problems the here presented method shall solve in space will now occur in the time domain.
Lee: Seasonal means are interesting when looking at inter-annual variability in nature, in the model and in the sampling bias. While not necessaries the best methodology for a pure validation exercise, looking at the seasonal mean of biogeochemical values allows us to justify the use of our model to inform policy.
Also, the same linear regression is also presented for the 'Full' dataset which has not been averaged, apart from in the case mentioned above where multiple in situ measurements were taken the same day at the same depths in the same latitude longitude pixel.

A better description of the difference between time granularity and time selection was added to the paper:

It is important to highlight the difference between time selection and time granularity. Once the matching was performed, both model and in situ datasets were studied under two time granularities: the annual mean and the daily mean, and under five different time selections: Annual, Winter, Spring, Summer and Autumn. All ten permutations of the two granularities and the five time selections were studied.

The time granularity defines how that data is aggregated, if at all. The daily or ``full'' granularity refers to a dataset containing every matched pair of points in the North Sea and the ``annual'' granularity is a series of annual means of that dataset. Typically, the full dataset contains many thousand matched pairs of in situ and model data, whereas the annual mean datasets contain only 35 points; one for every year between 1970 and 2005. The annual time granularity allows the study of inter-annual variability in nature, in the model and in the sampling bias. The ``full'' granularity allows the residual to be calculated for each in situ measurement, and is instrumental is the demonstration of the point to point matching.

The time selection specifies which part of the year is studied. As the AMM is a Northern Hemisphere domain, the Winter time selection masked out all measurements that did not occur in January, February or March. Similarly, the Spring contains the data from April, May and June, the Summer contains the data from July, August and September and the Autumn contains the data from October, November and December. The annual time selection effectively means that

no time selection was made and that data from all times of the year can be used.

As a shorthand, each combination of time selection and granularity can be described using the time granularity followed by the selection; for instance: Full Winter, or Annual Spring. The specific case of the annual granularity and annual selection is called the annual mean.

Lee: Previously, we had included only contain a subsection of those data:

- Annual mean **Winter** time selection,
- Annual mean **Spring** time selection,
- Annual mean **Summer** time selection,
- Annual mean **Autumn** time selection,
- **Annual** mean No time selection,
- **Full** time granularity No time selection .

These datasets were chosen because they cover a range of data that is understandable and allow us to demonstrate the value of the matching method. We now show the complete results of all ten permutations of granularity and selection, but we would like to move those tables to an appendix.

AR#1: Sec 4.2 (2317) line 19 and Sec 5: The standard error and the two tailed propability are not further considered in the text. Which information does one get from the change in these numbers?
Lee: As suggested by AR#2, we are in the process of reworking the entire Section 5, and this change is being implemented. The p value is the probability that these data are not derived by chance. The p-values are not a measure of goodness of fit; they are a measure of the confidence in the regression output parameters are correct, under the assumption that all data points have an equal influence on the fit. Given that the matching of dataset leads to some relationship between the in situ and the model, p-value tends to decrease with large sample size, even when the relationship is weak or non-linear. For this reason, we tend not to discuss the usefulness of p-values, but they are included for completeness.

AR#1: Sec 5 (2317) line 24: It would be nice to get an overview of the tables as graphs.The different regression lines could be plotted in one graph with different colors, for the seasonal consideration four short lines could be presented (see attached sketch). Generally, the content of the tables should be discussed in more detail.
Lee: As suggested by AR#1, the following plots were prepared from the results of the linear regressions. Each dataset is associated with two plots. The left hand shows the result of the linear regressions of the data with full granularity, whereas the right hand side plots show the linear regressions of the annual mean of the data.

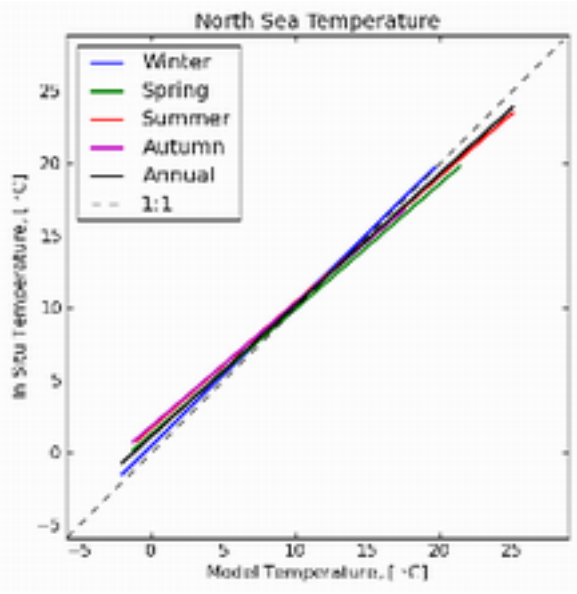
These plots allow us to look at the range of the dataset, and the Full dataset regression line extends far beyond the range of the annual mean. This makes sense as the “full” line is a linear

regression of tens of thousands of data, and the seasonal and annual linear regression each contain one point per year. It is valuable to show both, as it allows us to discuss the value of the point to point matching, using the “Full” datasets, but also allows us to

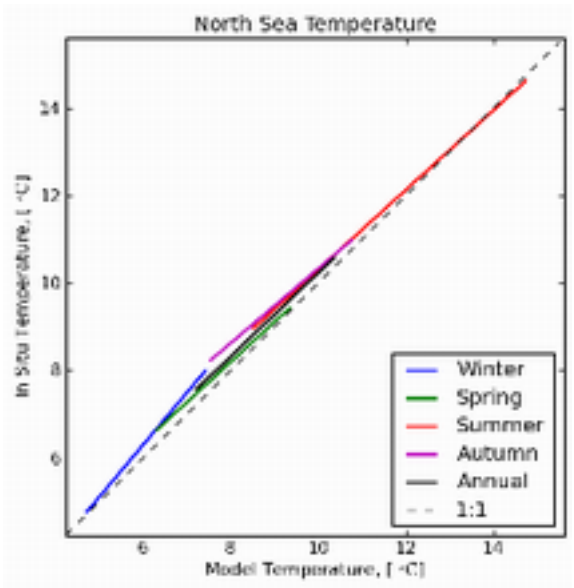
We are presenting these plots in the results section of this paper and we would like to move the tables 1 and 2 to an appendix.

Also, we’d like to thank AR#1 for this suggestion, it has led to a valuable new way to illustrate our model results:

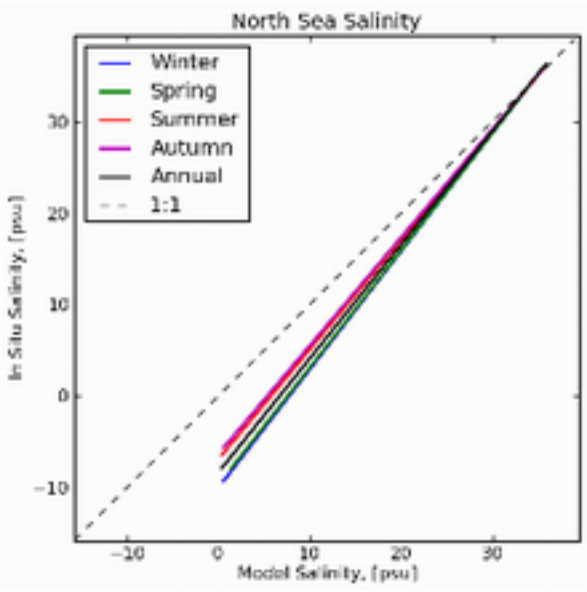
Full



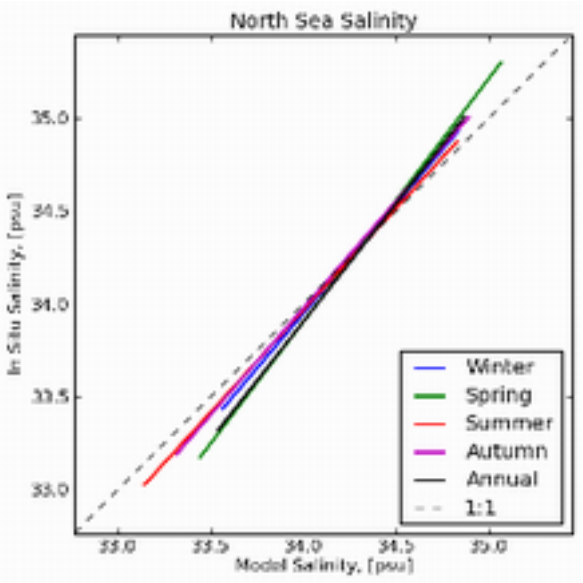
Annual Mean



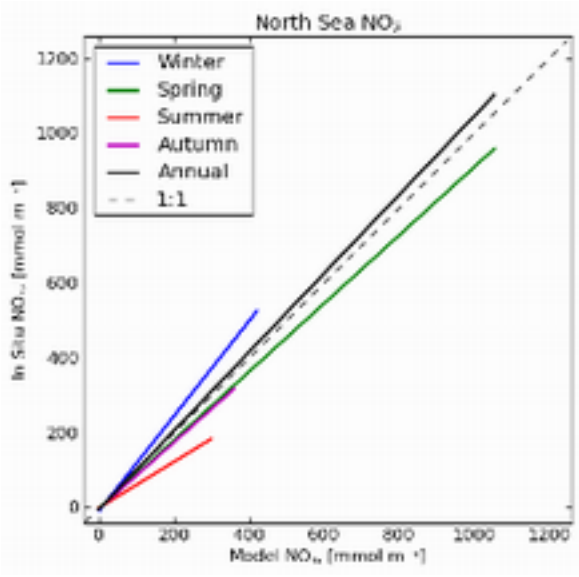
Full



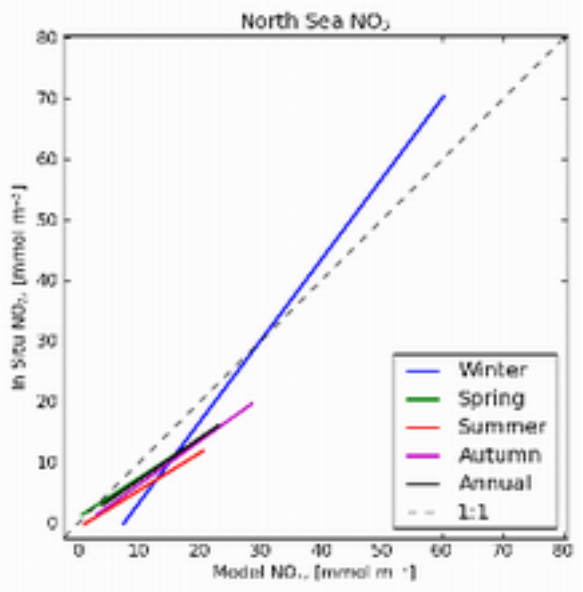
Annual Mean

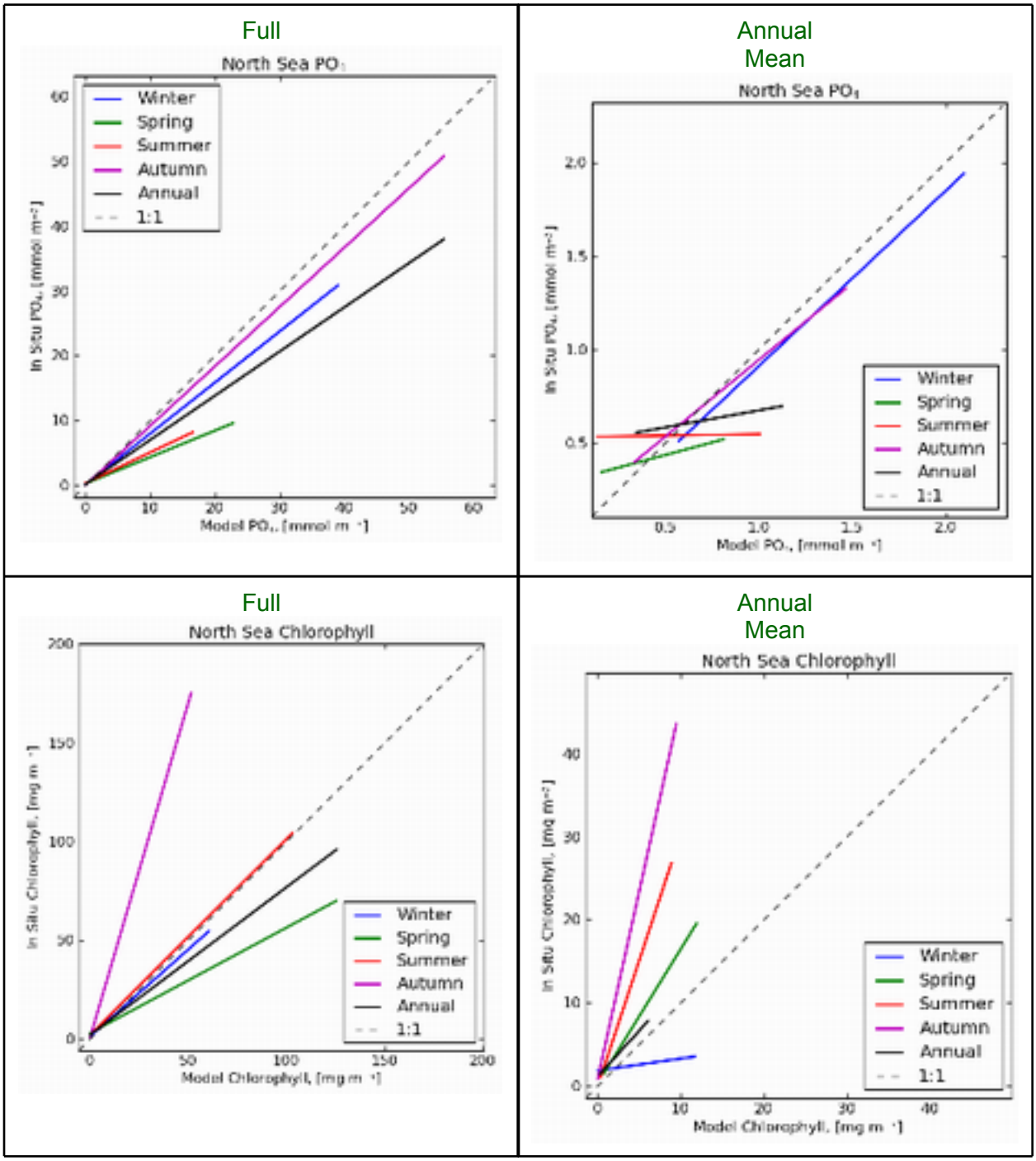


Full



Annual Mean





AR#1: Sec 5.1 (2318) line 6: The representation of figures 1-3 includes a redundancy. It is obvious that the regression lines fit best in the regions of high data density. This is an intrinsic feature of the linear regression.

Lee: While it is obvious that the regression line fit the region of high data density, it is interesting that the region of high data density coincides with the 1:1 line in these plots. This means that the

model is reproducing those measurements. The text has been changed to highlight this point. This would be clearer with non-logarithmic plots, perhaps.

AR#1: Sec 5.1 (2318) line 13: The reproduction seems to be extremly good because of the high number of data near the mean temperature. This only shows that the model does not produce a bias on temperature.
Lee: Text has been changed to: “The high density regions of these figures coincide with the line of slope unity through the origin. This shows that the model performed extremely well at reproducing the in situ salinity and temperature measurements.”

AR#1: Sec 5.1 (2318) line 24: Here the same problem occurs on the temporal scale the paper focus on on geographical scale. The time average equates here to the average over a large region. Thus, in addition to the here presented point-to-point method a time-to time method would improve the significance of the comparision.
Lee: This model it set up to save daily and monthly averages of various field and tracers, but the actual time step of the model is fifteen minutes. A work that presented a method to compare in situ data with the model time step instead of the daily mean was suggested in a short comment by H. Riede, (see above). They pointed out a paper that produced time step matched data at run time. As explained in the reply to their comment, this method was not possible for this work. However, it has been discussed in the revised text.
In terms of a time-to-time matching, we are matching the time as closely as is possible, given this model run. While it would be nice to be able to perform such a comparison, the typical range of variability of biogeochemical measurements over the course of a day doesn’t necessarily warrant higher temporal resolution, and the huge additional computing resources. At the moment, we are comparing a daily mean of the model to the instantaneous measurement, it can’t be more mismatched than comparing the mean of a small bottle of water to the mean of 144 square km of ocean.

AR#1: Figures related to Sec 5.1 (2331, 2332, 2333): The information of figures 1-3 is not clear in respect to the here suggested approach for comparing model results against in situ measurements. Understanding them as two dimensional histrograms of already matched model results versus in situ measurements, they could be a base for a discussion on the quality of the model itself but not to justify the here presented method.
Lee: The aim of this paper was to present this method and demonstrate that the model is a more or less accurate reflection of the historic Biogeochemical record, we feel that a discussion the quality of the model is a valid subject for discussions in this paper.

AR#1: Sec 5.2 (2319) line 3: Here two additional averaging processes have been included: annual mean and depth average. Again the question is how these averaging processes produce a bias and lead to misleading results.
Lee: The depth averaged data was included to illustrate both an example of a previous method to perform this study, but also was needed as a cross reference in order to test if there is some interannual changes in mean state of the modelled North Sea that would be hidden by the patchyness of the in situ dataset.

AR#1: Sec 5.2 (2321) line 15: It is argued that the quality of matching may depend on small difference in the timing of the bloom and thus, in the chlorophyll data. This show again the necessity of a time-to-time method. It should mentioned that according to section 2 the seasons are defined as JFM,AMJ,JAS,OND. This results in the fact that in some coastal parts of the ICES area the spring bloom may occur in the winter period.
Lee: When we argued that the there is a high sensitivity to bloom timing, the discrepancy would be of order days, instead of hours, having a higher temporal resolution would not necessarily improve the agreement of the current model and historical data, as the model is far from perfect at bloom timing.
In terms of having a hard border between the seasons, we changed the text: adding:
Additionally, the seasons were defined as strict three month periods, it is possible that for the earliest blooms, for instance in a coastal area, the Springtime bloom may have overflowed into the Winter bin.

AR#1: Figure 10 (2340): This figure gives the impression that the in situ data of chlorophyll have preferably been measured in high salinity regions over the years. The message of this figure is not clear within the context of this paper.
Lee: Figures 8 and 9 show a large increase in chlorophyll in the years 1990 onwards, but when we look at figure 10, it is clear that most of the chlorophyll measurements after 1990 were taken in low salinity coastal waters. Low salinity waters imply river influenced, high nutrient waters, so it's not surprising to find high quantities of chlorophyll producing phytoplankton there. However, the model had a coarseness of 12km, so can not resolve highly coastal bloom behaviour. Within the context of this paper, it is a demonstration of the use of this method to identify Model limitations.

AR#1: Sec 5.3 (2322): The here presented results of the ’target diagram’ method give confirmation of what the authors already have shown in the sections before: Spatial matching of model results to in situ measurements lead to better correlation of both when it comes to comparison and validation. When utilising target diagrams the results should be discussed in more detail, e. g.: why does matching apparently increase the correlation for the physical variables (temperature and salinity) but not for the ’biological’ variables (nitrate, phosphate, chlorophyll)? Is it a consequence of the matching model or a more or less singular result of this model. As mentioned this subsection should be combined with the subsections before.
Lee: The correlation increases for the "biological" variables too, but it is a less dramatic shift. Arguably, there is a larger apparently shift in the correlation of the physical elements of the model because they are modelled better. As you may expect, the Model is most successful at reproducing physics(Temp, Sal) , it is slightly worse at Chemistry (Nitrates, Phosphates) and even worse at Biology (Chlorophyll). Unfortunately, biology is much less tractable than physics. Also, Section headers 5.1, 5.2, 5.3 have been removed.

AR#1: Sec 6 (2323-2324) line 26 ff. - 2324 line 3: The third part of the conclusion somehow contradicts what the authors did state in the first and second part (2323 line 15 ff. and

2323 line 20 ff.). The matching of model results to in situ measurements clearly does not modifiy or improve the model.
Lee: Removed sentence: "Furthermore ...even after matching"

AR#1: Technical comments:
AR#1: Tables (2327-2330) and Figures (2334-2339) related to Sec 5.2: Table 1 and Table 2: Although mentioned earlier in the article, the two tables are missing the column for the autunm period. It is suggested to add the autumn column. Moreover, these two tables could be merged to one table or if this is not appropriate it would be better to have the more or less physical quantieties (T, Sal) in one table and the ecological in the other.
Lee: For completeness, we've now added the Autumn column. Typically, the autumn is an uninteresting season in Biogeochemistry, as the nutrients have mostly been used up, the plankton bloom is over, and the sea is frequently too rough for research cruises.
Secondly, the discussions format (GMDD) does not allow for large tables, but Tables 1-2 should be merged into one table in the GMD paper. For the meanwhile, I've moved Nitrates into the Table 2. We also would like to move these tables to an appendix.

AR#1: Table 3 and Table 4: Same as for table 1 and 2 (see above). Moreover, it would be helpful if the shown variables would be consistent. Fe, all variables evaluated as 'annual' or all variables evalutated for one season (summer, spring, etc.). It is difficult to estimate if here only the "best" results are shown or if the missing are "boring".
Lee: These combinations of variables and times (ie. Winter Nitrates)were chosen because of their value in informing policy and model validation. For instance, the annual nitrates (or even summer and Autumn Nitrates) are not biologically interesting, but the Winter Nitrates are very interesting.
As mentioned above, the discussions format (GMDD) does not allow for large tables, but Tables 3-4 should be merged into one table in the final GMD paper. For the meanwhile, I've moved Nitrates into the Table 4.

AR#1: Figures 4 - 9 (2334-2339) are recommended to be adapted accordingly. The figure captions should contain a cross refrence to the columns of the tables.
Lee: These figures are associated with tables 3 and 4, which will be merged into one table, and there will be a 1:1 correspondence between these figures and that table, which should be clearer.

AR#1: Sec 5.3 (2322) line 16: It is assumed that winter nitrate instead of winter nitrogen is meant.
Lee: Indeed, it was nitrates.

Reply to Anonymous Referee #2:
Anonymous Referee #2
Received and published: 14 December 2012

AR#2: Summary:
This paper demonstrates the difference between comparing numerical models to data by sub-sampling the model in time and space the same way as the in-situ data was collected as opposed to comparing spatial and temporal averages where all the model data have been used. This is demonstrated using a biogeochemical model and in-situ data from the North Sea. It demonstrates that some variability in the area-averaged mean of the in-situ data may arise from the sampling and if the model is not sampled the same way it may appear to not reproduce this 'variability'.
The method is just demonstrates for one particular model, not for models in general. The material is however presented in such a way that it should not be problematic to apply the model to another model/dataset, however that does not guarantee its success.

AR#2: Overall the paper i well written and the subject clearly presented, but I miss a more in-depth discussion of the results. I find the presentation of the method straight forward and easy to follow, however the result-section is mostly a description of the figures and not much discussion is provided. After that there is only a short conclusion.
I think the paper is appropriate for GMD and the method potentially very usefull to the earth science modeling community, but because of the shortcomings mentioned above I recommend this paper for publication only after a major revision has been performed. Below are more specific points for the authors to address.

AR#2: Specific comments:
Abstract:
"The application of the point-to-point method showed that the model was successful at reproducing interannual trends". Should it not be 'interannual variability'. If the variability only shows up because the system is sampled a certain way how can we know that this is actual interannual variability and not just a result of the way the system has been sampled? This is touched on this later in the paper.
Lee: changed "interannual trends" to "interannual variability". The idea of displaying the mean of the region in the model (in the grey area) allows us to compare the sampled data and the matched model against the unmatched model data. If there were some trend in the unmatched model data that isn't present in the matched data,

AR#2: "we advocate greater transparency in the publication of methodology.", I agree, but this is never mentioned again, so unless a paragraph about this is added it should not be in the abstract.
Lee: This sentence was removed from the abstract.

AR#2: Section 2:
AR#2: It says that the model is run from 1965 to 2004, why are the figures from 1970 only,

is this because there are no ICES data before 1970 or is there some other reason for excluding the first 5 years of the run? I thought the ERA40 reanalysis stopped in mid 2002, how come the run could go to 2004?
Lee: The model ran as a 45 year hindcast from january 1960 to december 2004, but the first ten years are spin up and not considered in the analysis. The ERA40 reanalysis was used for forcing the surface between 1960 and September 2001, subsequent conditions were from the ECMWF operational analysis. The “Model” section has been expanded and now describes the ERA 40 surface forcings and the freshwater fluxes in greater detail.

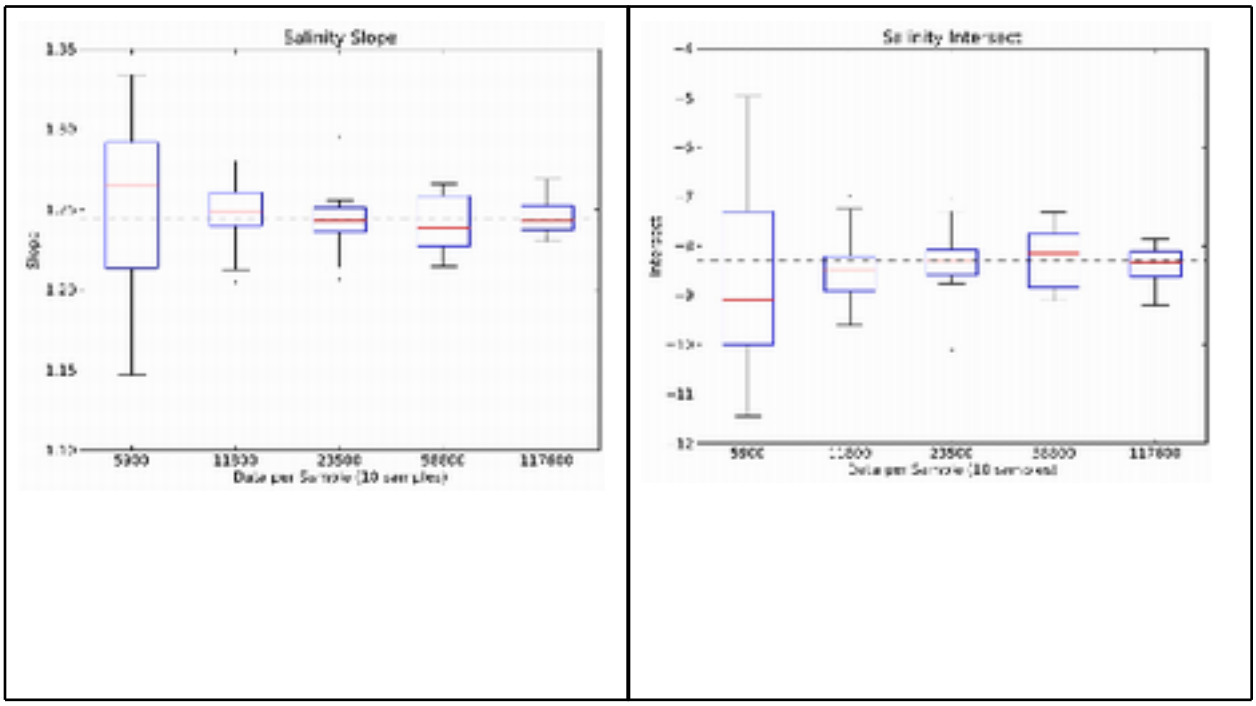
AR#2: Section 4:
AR#2: Page 2317, line 11: "These values ..." This sentence does not make sense to me, is something missing?
Lee: This sentence has been changed to:
The “entire region” mean is the mean of the North Sea model data before the point-to-point matching was applied. However, it is not reasonable to compare the mean of the entire water column to mixed layer in situ data as most measurements occurred in the mixed layer. For this reason, the entire region mean values were calculated as the monthly mean divided by the mixed layer depth.

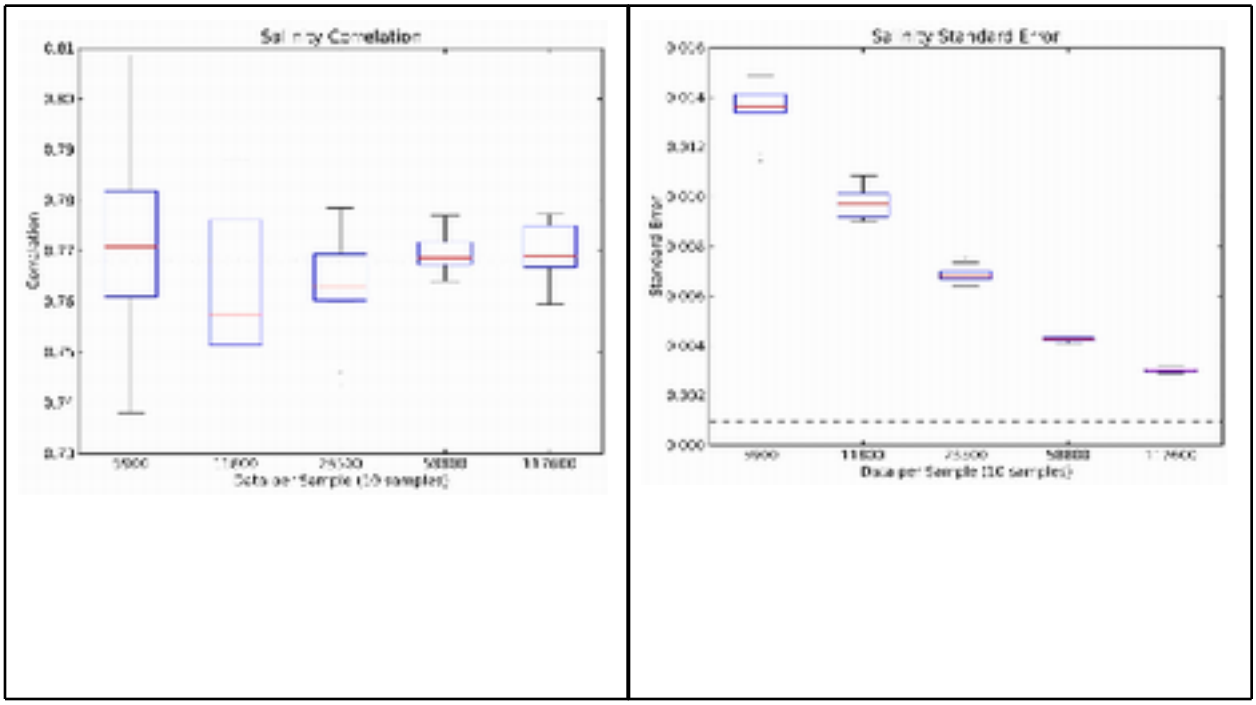
AR#2: Has any vertical interpolation of the model or in-situ data been done? What is the vertical resolution of the model compared to the in-situ data?
Lee: Due to the complex nature of the water columns vertical structure, it is extremely difficult to interpolate sparse marine in situ measurements at depth into a three dimensional grid.
In the case of a three dimensional interpolation of ocean measurements, the vertical scale has a much smaller scale of variability than the horizontal scale, complicated the weighting of data. Typically, 3D interpolation assumes that all axes have equal weight, but in the ocean, the vertical scale varies in distances of meters and the horizontal scales in kilometers. It is not straightforward to choose whether a pixel is more influenced by a deeper nearby point or by a distant point of equal depth.

AR#2: Section 5:
AR#2: In this section I miss more discussion of the methods and results - for example:
-There are many salinity and temperature mesurements and few nutrients and chlorophyll measurements - is there a difference in how the method perform in relation to how many measurements there are?
Lee: To address the question of the method performance against the number of measurements, we performed the following exercise, which could be added to the note:
Each of the five in situ measurement datasets and their point-to-point matched model datasets (Salinity, Temperature, Chl, n3, p4) in the North Sea were subsampled and then a linear regression was applied to the subsamples. The output parameters of the linear regression (slope, intersect, correlation, p-value, standard error) were plotted in "Box and Whisker" diagrams.

The subsampling was performed first by randomly selecting 10 independent samples of the total dataset. The selection of 10 independent samples was performed 5 times, and the size of the subsamples were varired between: 1/200th, 1/100th, 1/50th, 1/20th and 1/10th of the total dataset (rounded to the nearest 100 data points). i.e. if the total dataset had 200 000 data points, then the follow subsamples were prepared: 10 x 1000, 10 x 2000, 10x4000, 10 x 10000, 10 x 20000. The subsamples contained independent measurements to the other samples of the same size, but may overlap with subsamples of a different size.

These figures (below) show the slope, intersect, correlation and standard error for temperature. As the subsample of data becomes more representative of the total dataset, the results of its linear regression converge on the linear regression of the total dataset, shown by the black dashed horizontal line. However, adding more data does not make the linear regression converge of the ideal result.





This highlights is that the model skill does not scale with the number of in situ data. That is fairly obvious, as the number of in situ data is completely independent of the quality of the model. More importantly, this exercise demonstrates that the point-to-point matching method doesn't have any data quantity or distribution requirements.

AR#2: -In the introduction it is mentioned that when a region is well sampled, interpolation can also be used, but what qualifies as 'well sampled' and what is 'sparsely sampled'? Additionally at the end of section 5.2 it is mentioned that a more diverse and regular dataset of chlorophyll is needed to use the point-to-point method, but isn't this precisely the kind of dataset this method is supposed to accomodate?
Lee:
As mentioned above, if the typical distance between data points is larger than the scale of variability of the data, then interpolation would fail to produce a meaningful dataset. For this reason, interpolation works best in regions with low variability or high data occupancy. In theory, the interpolated dataset could be used with some measure of uncertainty related to how well the region was sampled. However, in our case, it would be a needlessly complicated extension of the matching method. Additionally, in the case of a three dimensional interpolation of ocean measurements, the vertical scale has a much smaller scale of variability than the horizontal scale, complicated the weighting of data. Typically, 3D interpolation assumes that all axes have equal weight, but in the ocean, the vertical scale varies in distances of meters and the horizontal scales in kilometers. It is not straightforward to choose whether a pixel is more influenced by a deeper nearby point or by a distant point of equal depth.

For the second point here, I agree: we don't necessarily need a more diverse data set, but rather a better chlorophyll model. I have rewritten this paragraph to emphasize the limitations of the Biogeochemical model, instead of the dataset:
Although much of the in situ variability of the larger datasets (temperature, salinity, nitrates, and phosphates) can be accounted for, the model does not reproduce many of the historic trends of the in situ chlorophyll measurements on a point-to-point basis. While a more diverse distribution of North Sea chlorophyll measurements would help to validate the model, it's not possible to travel back in time to obtain such a dataset. The failure to reproduce the historic data time series may be due to the effects of sub-pixel variability, in which case higher resolution models could allow a point-to-point study of chlorophyll to converge on the in situ measurements. It is also possible that a better match between the in situ measurements and the model may yet occur through improvements to the phytoplankton model.

AR#2: -Is there a lower limit to how sparse the dataset can be in order to use the point-to-point method?
Lee: The point to point method can be applied with any number of data, as demonstrated above.
There's a theoretical limit of two data points needed to perform a linear regression, but that is not a practical limit. The linear regression has smaller standard error with larger datasets.
In terms of technical limitations, we found that performing the matching becomes memory intensive with the larger datasets.
If the goal of the study is to identify long term interannual variability, then there needs to be sufficient data in each year to identify trends.

AR#2: -How does the spatial and temporal sampling bias affects the calculated means?
Lee: Spatial sampling bias can make a large change of the mean. In figure 8, 9 and 10, the chlorophyll in the Spring and Summer appears to be extremely high, making the in situ data appear to have a very large bloom that is not captured by the model. However, when we look at the salinity of the samples taken, we see the chlorophyll measurements after 1995 were almost entirely measured in low salinity (River influenced) waters. We effectively had a river-biased dataset for those years, and the model is not designed to be able to deal with those effects, so the model appears worse than it is. However, due to the point to point matching, we capture some of the trends, but not the amplitude of the bias induced spikes in measured chlorophyll.

AR#2: -For which other types of models may the method be useful?
Lee: Both the point-to-point matching method and linear regression are already used by the modelling community. It would be best put to use in situations where the quantity of in situ data is very low relative to the number of three dimensional pixels in the model.

AR#2: -What challenges may other models face when using this method for model comparison?

Lee: This method is not ideal in situations where the model pixel size is much greater than the scale of variability of the measurements. To use a marine example, at the mouth of a river, the salinity may go from nearly 0psu near the river mouth to 35psu 10km away in the sea. However, this model has 12km x 12km pixel size, and all those data between the river mouth and 10km offshore falls into the same model pixel. As such, point-to-point matching is not ideal for studying regions with high sub-pixel variability.

AR#2: I find it a bit odd that the sub-chapter are named after the figure-type rather than subject.
Lee: These subchapters were removed as requested by AR#1.

AR#2: Page 2318, line 13: "The high density ... " I agree that from the density plots temperature are well reproduced by the model, but I cannot agree with that for salinity.
Lee: This was also noticed by AR#1 and has been re-worded appropriately. Text has been changed to: “The high density regions of these figures coincide with the line of slope unity through the origin. This shows that the model performed extremely well at reproducing the in situ salinity and temperature measurements.”

AR#2: Page 2319: Are the modeled sea surface temperatures actually relaxed/nudged towards SSTs from the ERA40 dataset or just forced with ERA40 heat and momentum fluxes? In any case it seems the model does a good job with the temperature, I think it is unnecessary to give ERA40 so much credit for this. Perhaps rather add a sentence or two about how the SST forcing is done in the model description.
Lee: I’ve added a fuller description of the surface forcing, and changed the text to allow ERSEM to take more credit:
As the model surface temperature is forced using reanalysis based on aggregated observational data, it is not surprising that the temperature section of table X shows a strong correlation between the model and the in situ data. However, the atmospheric forcing dataset, ERA40-reanalysis, is a meteorological surface dataset, whereas the in situ measurements and hence the matched model data may occur at any depth. As such, the success of the model is due to its own merit, instead of the similarities between the forcing and in situ datasets.

AR#2: Tables:
AR#2: Is there any particular reason why autumn not included in the tables? I don’t understand what is the difference between ’Full’ and ’Annual’, they have exactly the same N, but somehow get different linear regression parameters.
Lee: The Autumn column was also requested by AR#1 and has now been added to these tables. As we mentionned above, it was left out because the autumn is an uninteresting season in Biogeochemistry, as the nutrients have mostly been used up, the plankton bloom is over.

The ``full'' granularity compared every matched pair of points in the given region, the ``seasonal'' granularities were the mean of the data in three month blocks, and the ``annual''

granularity is the annual mean of the data.The “annual” and the ``full'' columns contain the same number of data, but the annual column is the result of a linear regression with one point per year, whereas the ``full'' column is a linear regression of the entire point-to-point dataset.

AR#2: Figures:
AR#2: It would be helpfull to see a map of both the model region and the ICES subdivision IV. When variables are shown from the entire model in the figures , is it the entire model or just the all the model data in ICES subdivision IV?
Lee: The plot below has been prepared to illustrate the boundaries of the model domains. When data is described as the “Entire model”, it is the dataset in the North Sea before any data were masked by the point-to-point matching method. Region others than the North Sea were not studied in this paper.

AR#2: I don’t understand what the gray area in the figures 4-9 represents, what is the meaning of plotting the area and not just a curve? In the text it says that is the value averaged over the mixed layer, if so, this should also also be what it says in the figure label, not ’depth integrated’.
Lee: It was plotted as an area instead of a curve simply to keep some clarity in a greyscale plot. I’ve (hopefully) clarified the description of the grey area at the beginning of the time series subsection:
These figures each contain three curves: the matched model data (black line), the in situ data (dotted line), and the mixed layer depth-average of the entire model region (grey area). The

entire model region plots (grey area) is the model data in the North Sea region before any data were masked by the point-to-point matching method. This is included to estimate whether the matched and in situ variation correspond to overall trends, or sampling biases. The depth averaged data is included to illustrate both an example of a previous way to perform this study, but also as a cross reference in order to test if there is some inter-annual changes in mean state of the North Sea in the model that could be hidden by the patchiness of the in situ dataset.

AR#2: Technical details/suggestions
Page 2313, line 16: delete ’of’
Page 2316, line18: "forced into the same grid as the model " it sounds a bit like the data points are being moved, would it be better to say "collected according to grid cells" for example.
page 2317, line 8; delete "as the AMM is the Northern Hemisphere domain" and add "defined as" after "was".
page 2318, line 4: Whys do nitrates and phosphates have the subscripts 3 and4?
page 2318, line 7: ’nitrates’ should be ’nutrients’ I guess.
page 2318, line 20-21: remove the ’d’ at the end of ’overestimated’ and ’underestimated’.
page 2318 line 20: instead of using "extreme" indicate if it is fresh/salt water or high/low nutrient values the model misses.
page 2319, line 19: "larger" should be "higher temperature".
page 2321, line 12. Find a better word than ’performant’. As far as I understand performant means efficient.
page 2323, line 10: delete ’simple’.
page 2324, line 7: ’more large and long-term’=> ’larger and longer ’.
page 2324, line 5: delete ’hidden’.
Lee: These have all been implemented, except p2318, line 7: we did mean nitrates, not nutrients.