Review 1

Gary Brassington

The authors would like to thank Gary for his review of this work and for taking the time to read through the paper in such detail.

Abstract

The FOAM system is described in the abstract and early section as inclusive of multiple models of varying resolutions. However the validation only focuses on the global system. As the focus is on the performance of the global system I suggest removing mention of the other systems from the abstract and retain in the introduction.

The paper itself is quite large and so we are conscious of the fact that to include everything would make it unmanageable. This is the reason that the regional models are not introduced fully and assessed in the paper.

We agree that having mention of them in the abstract is misleading and have therefore removed mention of them from the abstract whilst keeping the brief mention in the introduction as you suggest.

The v12 system includes many significant changes. The control experiment chosen was to perform a free running model. It might have been more instructive to perform a data assimilative run with fewer of the changes to provide some comparison on what changes might be responsible for the positive and negative results.

Owing to the volume of information involved it has not been possible to present details of all of the R&D work that has gone into the creation of this system.

The following trials were performed in addition to those presented in this work:

1. bulk forcing vs. direct fluxes SBC using FOAM v12 system which show that the use of bulk formulae have a positive effect almost everywhere and in particular for the sea ice
2. NEMOVAR vs. OCNASM using FOAM v12 system which can be found detailed in Waters et al. 2014
3. LIM2 vs. CICE ice models using FOAM v11 system which show a significant improvement in the ice concentration fields.

The results from these trials have been used to support some of the observations and discussion in Sections 4 and 5.

Section 2.1:

The instability is a numerical one....The problem is not just due to unrealistically high diffusion.

The comment about horizontal scaling of diffusion coefficient has been modified to explicitly mention 'numerical instability'.

Section 2.2:
Are only OGDR's used for altimetry, what data volume and coverage is assimilated on average from each platform in each of the two 24 hr analysis windows.
Yes only OGDRs are used for altimetry both operationally and for the reanalysis trials.
At present, and after QC has been performed, the operational FOAM system is using approx. 60k observations per day during the (T-48h,T-24h] analyses and 40k observations per day during the (T-24h,T+0h] analyses on average.
Coverage of tracks is global even during (T-24h,T+0h] (an example image for the observations used for the (T-24h,T+0h] analysis on 25th February this year can be found in Figure 1 of the author response gmdd-6-C2965-2014.pdf) although, of course, there are gaps between tracks which are filled approximately every 12-15 days.
Obviously data volume and coverage for SLA differ significantly depending on the state of the satellite observing system.
However the statistics presented for the the reanalysis trials in Section 4.1 are calculated using an average of 47.5k observations per day (after QC & filtering to a common subset).

Section 2.3:
Why is T-54h required for the atmospheric forcing?
The `T-54h' in bullet 2. of Section 2.3 was a typo and has been changed to `T-48h'.

What is the resolution of the atmospheric model used in the hindcasts? Could you add this to Table 1
The NWP forcing fields used throughout the entire reanalysis trial period are from the UM Global system introduced in Section 2 which runs at approx. 25km resolution. Given that this is the case for all the reanalyses and the operational system that the paper is describing it would seem unnecessary to add it to the table.
However the paper does not make it clear that the forcing for all 3 trials came from the same NWP model running at the same resolution throughout the trial period and so I have added a sentence to the start of Section 3 to explain this.

48-hour assimilation window
There seems to be some confusion relating to the 48-hour window that is used operationally and the 24-hour daily cycling used in the reanalyses.
Given that Reviewer #2 also had some questions relating to this it is evident that the explanation provided needs to be made clearer.

When we run FOAM operationally in near-real-time we start from T-48h and produce our `best estimate' analysis for the (T-48h,T-24h] period.
The model state at T-24h is then saved for the next day as our best estimate of the ocean state. An update run is then performed for the period (T-24h,T+00h] before we start the 7-day forecast.
There is no doubling up or extra weight applied to any observations - we have simply

moved our `best estimate' back by 24-hours by running an analysis for the period (T-48h,T-24h].
The impact of running a second analysis period therefore is to allow more observations to be used in the generation of our best estimate initial conditions each day owing to the late arrival of some observations.
As mentioned in Section 2.4.2 (now Appendix A) this led to a RMS error reduction of approx 5-6% globally in the operational near-real-time system when it was implemented.
To make this clearer in the paper I have modified the explanation of the 48-hour window in Section 2.3 to better distinguish between the `best estimate' analysis and the `update run'.

We do not use this approach for the reanalyses because they are run in delayed time (typically > 6 months after real-time) and so have access to more observations than the operational system because late arrivals will be included.
These reanalyses are not an identical copy of the operational suite and are used in a calibration sense to understand the potential errors in the operational system.

I have therefore expanded the last sentence at the end of the 1st paragraph of Section 3 to explain that we do not need to run 2 separate assimilation cycles in the reanalyses because the observations are extracted in delayed-time and are therefore more plentiful than those used operationally.

Does Water et al., 2013 show whether FGAT matters over a 24h time window?
Waters et al. 2014 do not investigate what impact the use of FGAT has on the FOAM system.

"reanalysis observations are filtered. . ." do you mean sorted to a common subset?
Yes I do. The text has been changed to explicitly say `common subset'.

Bar-chart figures
The scale for the plots is being set by the free run.... I suggest that the x-axis be reduced and where the free model goes beyond the limit place a bracketed value above the line to indicate its value.
Figures 1 & 2 have been re-worked according to your suggestions which have improved the readability of the figures.
In all plots where the maximum error for either of the v12/v22 runs is less than 2/3 of the free run, the x-axis scale is truncated so the finer detail can been seen for the assimilative runs.
Where this happens the error value for the free run is annotated to the plot above and at the end of the corresponding error bar.
Those plots that have changed can be found in Figures 1 and 2 in the updated paper and in Figure 2 of the author response gmdd-6-C2965-2014.pdf.

Profile error figures

It would be instructive to see the tropical Pacific as an additional 2 panels to shed further light on what the distribution of error is in this region.
Figure 3 has been extended to include temperature and salinity error profiles for the Tropical Pacific and the North Atlantic. The figure is now also larger to make the detail more visible. This can be seen in the updated paper or in Figure 3 of the author response gmdd-6-C2965-2014.pdf.

Salinity profile errors
The paper indicates that there is a general deterioration in the performance of v12 in the ocean interior. This requires some further discussion.
A comment has been added to Section 4.1.3 to explain the general deterioration in the salinity fields in the ocean interior at v12.
This is related to the difficulties fitting the relatively sparse salinity observations with short correlation length-scales and it is hoped that the adoption of dual length-scales will improve things in the ocean interior.

Previous FOAM systems have converted altimetry into synthetic profiles. Is this still the case?
FOAM does not convert altimeter observations into synthetic profiles.

The large bias in the North Atlantic is associated with a large cool bias in the temperature. Whilst the Mediterranean is associated with a warm bias. Is this consistent with the precipitation hypothesis?
The text has been changed from 'believed to be caused by excessively high precipitation in the surface forcing fields' to 'believed to be an artefact of the increased number of coastal observations in these areas'.
The original text was a little misleading and suggested that precipitation biases are worse in these areas which is not necessarily the case. Rather these biases have more impact in these regions owing to the number of coastal salinity observations in these regions.

NEMO is a volume conserving model. The long term drift in the model needs to be explained by a change in volume rather than one in terms of steric expansion such as low density water. A volume conserving model does adjust sea level for steric effects but this is done without any net change to the global volume.
Yes this was misleading and has been changed to simply explain that the SSH increase is caused by a mismatch between the precipitation and riverine freshwater inputs.

Near-surface velocities
It is worth emphasising that this verification is based on independent observations.
A sentence has been added to the 1st paragraph of Section 4.1.6 to emphasise that the velocity validation is based on independent data.

How many observations are being used in each basin?
Re. data volume of drifter-derived velocity observations approx. 725 drifter observations per day are used globally for the assessments in Section 4.1.6.
The 1st paragraph of Section 4.1.6 has been modified to include this number.
For your information the regional breakdown is roughly ~300 per day in each of the Atlantic and Pacific Oceans, ~65 per day in the Indian Ocean and ~150 per day in the Southern Ocean. (NB. these figures sum to more than 725 because the Southern Ocean domain overlaps with each of the other 3 to some extent.)

Why is the Indian Ocean excluded?
I am not sure I follow your comment re. the exclusion of the Indian Ocean. The Indian Ocean is included in the velocity validation and is mentioned in the text (Section 4.1.6). In particular it is noted that v12 velocities are better in the Indian Ocean in contrast to the Tropical Pacific and Tropical Atlantic where v11 is better.
If the question is why is it not included in the Taylor plots in Figure 5 then this is because I think that 4 regions (with 3 points per region) is the most that can fit on these Taylor plots without making them too complicated.
As described in the text, the results were generally consistent across the regions – being better in all extra-tropical regions and worse in the Tropical Pacific & Tropical Atlantic.
It was therefore decided that we would only show results from the global ocean and 1 region for each of the tropical, mid-latitude, and high-latitude regimes (namely Global, Tropical Pacific, North Atlantic & Southern Ocean) to illustrate this.
Results for all the 8 regions (apart from global) can be found in Figure 7 of the author response gmdd-6-C2965-2014.pdf for your interest.

For the Tropical Pacific both the meridional and zonal components seem to be inferior to v11 which is not consistent with the text.
This is explicitly mentioned in the text of Section 4.1.6 as follows:
"Although better in the Indian Ocean the v12 system is worse elsewhere in the tropics; in particular in the Tropical Pacific. Further comparisons with currents measured by the TAO/TRITON (McPhaden et al., 1998) and PIRATA (Servain et al., 1998) tropical moorings (not shown) confirm the findings of the drifter regional results that the skill of current predictions is reduced in the Tropical Pacific and Tropical Atlantic.

Is this the DA u,v fields or the initialisation shock of the T/S/eta state? Do you track KE during the initialisation?
This is the assimilation of T/S/SLA not the velocity balancing. The latter is geostrophic and therefore effectively zero near the equator.
In the text I have replaced 'increments' with 'tracer increments' to make this clearer.
We do not track KE through the IAU as standard but have performed a few quick runs to check and can see no evidence of shock.

Increase fontsize – difficult to read in this scaled down version. Also use the full

column width.
I have increased the size of the Taylor plots in Figure 5 as per your suggestion. Please see the end of this response regarding font sizes.

It is instructive to see errors in each component. However, it is also useful to perform the analysis on the total vector. For example, Kundu, 1976, JPO using a complex correlation.
Although the Kundu and Allen (1976) technique looks potentially interesting we would not be comfortable including this at this stage. This will be considered as a useful extension for future validation exercises.

Forecast validation
It would be instructive to compare the power spectrum of the analysis and the forecasts. Is there larger power in the high wavenumbers as speculated which is subsequently dissipated through the forecast period?
Yes this would be instructive and we plan to perform this sort of analysis in the future to guide development.

It is worth noting that the hypothesis of over mixing for the temperature biases does not seem to be present in the salinity results.
A comment has been added to Section 4.2.3 to explain that, although global salinity profiles do not show evidence of excessive mixing, the mixing bias is present in mid-latitude regions. An example for the North Atlantic can be found attached Figure 4 of the author response gmdd-6-C2965-2014.pdf.

Comparisons with gridded observations (Section 4.3)
The example shown would benefit from the calculation of spatial correlations and included in the text to quantify the improvement of v12.
Yes we agree! Anomaly correlation have been calculated for the 2D fields shown in Fig. 10 over the Agulhas retroflection region which can be found in the new Table 4. These results support the conclusions of the qualitative assessments described in Section 4.3 that the v12 system provides a better representation of surface mesoscale fields.
Subsequent calculation of spatial correlations for other case studies confirm the statement that, in general, the v12 fields show a better agreement with the observations.

There is a lot of material introduced for a single case study example. If supplementary material is permitted it would be desirable to show at least two other examples.
The Agulhas case study was chosen because there were some fairly obvious and interesting features present at this particular time.
I have included a figure and table analogous to Fig 10 & Table 4 for a second case study covering the East Australia Current region which can be found in the supplementary pdf file.

A further extension to this (as suggested by Reviewer #2) has been to quantify the relative improvement to the mesoscale eddy fields in the new system by assessing the SLA and near-surface current fields separately for areas of high and low mesoscale variability.
These high/low variability regions were defined using the spatial variability distribution of SLA observations over the 2 year assessment period by using a threshold standard deviation.
These results can be seen in the new Table 3 and are discussed in Section 4.1.4.
Results show that the improvement at v12 in areas of high variability is considerably more pronounced than for areas of low variability which is consistent with the findings presented in Section 4.3 and discussed throughout the paper.

Summary
The opening statement of the 2nd paragraph must state that the results are mixed. There are clear advantages when the observation density is high but for regions with sparse observations the performance has deteriorated.
Paragraph 2 of the Summary (Section 5) has been modified to say that results are mixed with considerable improvement where observations density is high but with some deterioration in areas of sparse observations.

There is no information/diagnostics presented on the initialisation shock such as global KE. Given the rapid deterioration in the T/S profiles in the forecast compared with v11 these diagnostics would be instructive.
We have not included any kinetic energy diagnostics in the paper but have performed a few short tests to output global KE at each model time step.
These tests show no sign of any shock and no suggestion that the situation is different at v12 than at v11. So we are confident that the IAU is performing as expected.
What is shown however is that the data assimilation is increasing the KE in the system (or rather preventing the global KE from decreasing).
This is expected given that the 1/4 degree model is not eddy resolving and relies on the DA to provide some of the eddy variability.

The rapid deterioration in profile error is a perhaps a little exaggerated by the comparisons between forecasts and analyses.
By comparing with the analysis daily-mean fields we are essentially comparing against a field where the observation in question has already been assimilated (certainly approx. half of the increments have been applied).
Whereas for forecast day 1 the system will not have assimilated data from this instrument for a number of days (10 for Argo) and so the deterioration will have happened over the 10 days not 1 day and in that respect the lead-time plots.
So this is partly owing to over-fitting but also to the sparsity of sub-surface observations (in time and space).

Figure font sizes
Reviewer #2 also commented on the relatively small font size used for the figures and so we acknowledge that this will need to be increased.

The font sizes used in the figures are either 18 or 20pt but it is the figure scaling employed by the typesetting that causes them to be smaller than the journal text. Ideally these should align with the font-size used for the figure caption.
When the final typesetting is done we shall make sure, in conjunction with the journal, that the fonts used in the figures are clear and in keeping with GMD guidelines.

We have not done so as part of this response because the figure scaling used for this (single column) GMDD discussion document will most likely be different from that used in the (dual column) final version – meaning that any changes we made now may very well need to be redone at the typesetting stage.

Review 2

Anonymous

The authors would like to thank Reviewer #2 for a their review of this work and for taking the time to read through the paper in such detail.

Use of 48-hour assimilation window

There seems to be some confusion relating to the 48-hour window that is used operationally and the 24-hour daily cycling used in the reanalyses.
Given that Reviewer #1 also had some questions relating to this it is evident that the explanation provided needs to be made clearer.

When we run FOAM operationally in near-real-time we start from T-48h and produce our `best estimate' analysis for the (T-48h,T-24h] period.
The model state at T-24h is then saved for the next day as our best estimate of the ocean state. An update run is then performed for the period (T-24h,T+00h] before we start the 7-day forecast.
The impact of running a second analysis period therefore is to allow more observations to be used in the generation of out best-guess initial conditions each day owing to the late arrival of some observations. As mentioned in Section 2.4.2 (now Appendix A) this led to a RMS error reduction of approx 5-6\% globally in the operational near-real-time system when it was implemented.
To make this clearer in the paper I have modified the explanation of the 48-hour window in Section 2.3 to better distinguish between the `best estimate' analysis and the `update run'.

We do not use this approach for the reanalyses because they are run in delayed time (typically > 6 months after real-time) and so have access to more observations than the operational system because late arrivals will be included.
These reanalyses are not an identical copy of the operational suite and are used in a calibration sense to understand the potential errors in the operational system.

I have therefore expanded the last sentence at the end of the 1st paragraph of Section 3 to explain that we do not need to run 2 separate assimilation cycles in the reanalyses because the observations are extracted in delayed-time and are therefore more plentiful than those used operationally.

Re-structure of Section 2

Details of the v10 -> v11 upgrade have now been moved into an appendix (Appendix A) and Section 2.4 has been simplified.
However we do not feel that the NEMO/NEMOVAR details in Section 2.1 and 2.2 should be moved to appendices because we think, as this is a system description, these are an important focus of the paper.

Collins et al. reference

I don't think Collins et al. 2006 support the affirmation that "the Atlantic meridional overturning circulation at 26.5 N" is "important for the initialisation of the coupled seasonal forecasts". Its focus is on interannual to decadal forecasts.

Yes you are correct this is more aimed at inter-annual to decadal.

I was trying to say 2 things here; 1. that the improved representation of the mesoscale is beneficial when we use FOAM to initialise our coupled seasonal forecasts and 2. that an improved representation of the AMOC leads to improvements on longer time-scales.

This sentence has been modified and the Collins et al. reference replaced with 2 additional references: Barnston et al., (2012) who discuss the importance of an improved initialisation for seasonal forecasts and Cunningham et al. (2013) who present some recent observational results demonstrating the potential importance of the AMOC in controlling sub-surface temperature anomalies in the sub-tropical Atlantic.

Sea ice

Could the authors be more specific? What are the inconstancies implied by LIM2?

The inconsistencies implied by using LIM2 are simply that it is not the same model as is used in the other Met Office forecasting systems (seasonal, decadal and climate) which use the CICE sea ice model with 5 thickness categories.

The goal of the Met Office is to develop a consistent, seamless approach to forecasting across all time scales which is described further by way of the addition of Brown et al., (2012) to the references.

Aside from the seamless agenda, consistency is particularly important for the GloSea5 seasonal forecasting system which is initialised using FOAM ice analyses each day. We therefore require FOAM and GloSea to be as consistent as possible to reduce the chance of coupled initialisation shock.

It is not clear whether there is some balance relationship between sea-ice concentration and the other state variables (none of that is in Weaver et al, 2005). Could the authors clarify this?

Sea ice concentration has been implemented as an unbalanced variable in the linearised balance relationships and so is not balanced with respect to the other state variables. This is described in Walters et al. (2014) and I have added a sentence to Section 2.2 to make this clearer.

Is there a constraint that the ice thickness is positive within the assimilation scheme, or is this ensured by the model?

Although the model would prevent this happening the assimilation does not actually make any reduction to the category mean ice thickness and so this is not actually possible.

Persistent warm bias at 100m

The formulation "NEMOVAR fails to fully constrain a persistent model bias" is a bit

specious... this implicitly says that OCNASM succeeded reducing it in v11. What is the bias of the v11 equivalent free run?
This is a persistent bias because, as well as being present in the free-running NEMO model at v12, it is also apparent in the v11 free run (not shown).
The difference between v12 and v11 is that the NEMOVAR assimilation has not managed to constrain this bias as effectively as the old OCNASM system.

This issue is related to the difficulties associated with fitting the relatively sparse sub-surface observations with the short correlation length-scales employed by NEMOVAR.
It is hoped that the adoption of dual length-scales will improve things in the ocean interior and development of this is underway.
This issue is mentioned at the end of Section 4.1.3 and in the Summary.

NEMOVAR constraining mesoscale eddies better
It is claimed several times in the paper that the NEMOVAR assimilation scheme is more suitable for constraining the eddy variability, but it is not fully demonstrated in the paper in my opinion.
The evidence for these statements is based upon the gridded data comparisons in Section 4.3 supported by the findings of Waters et al. (2014).

However we are inclined to agree with you when you say you say that you don't feel that this has been fully demonstrated. The approach that you suggest seems very sensible and so we have adopted this to add some clarity to this issue.
To do this the extra-tropical ocean (between 23-66 latitude) was partitioned into high and low variability regimes based on the standard deviation of SLA observations for the full 2-year assessment period.
Statistics for SLA and near-surface currents were then calculated for each of these high/low variability regimes for both the v11 and v12 systems.
We then calculated the relative improvement (as a percentage reduction in RMS error) of v12 over v11 for each of variables and regimes.

Results show that the improvement in SLA for v12 is a factor of 10 higher in high variability regimes than for low variability regimes.
The same is true for near-surface velocities which show a 2-fold reduction in RMS error in high variability compared with low variability regions.

These percentage improvements can be found in the new Table 3 and are the process is introduced and discussed in Section 4.1.4

This section does not present forecast statistics of sea level anomalies. I wonder why, because sea level is a useful indicator of the upper ocean dynamics.
Sea level anomaly forecast statistics have not been calculated because, owing to a problem with our archiving, we do not have all the data required to perform the validation. Unfortunately the altimeter bias file used each day to calculate the SSH

from the SLA was not archived correctly.
The only way to generate these statistics would be to re-run the trials in their entirety, which is not possible at present. This issue has been corrected and so these results will be included in future work.

The problem of overfitting has already been addressed within the NEMOVAR system in Daget et al., 2009. They propose a diagnostic that could be used here.
The error variances will be recalculated as part of dual length-scale changes and this will be a useful metric to diagnose over-fitting as part of this change.

Remark about Figure 7 b and d: the forecast lead-time information is not obvious to understand. I suggest a better caption text, or a different representation.
The caption of Fig. 7 has been modified to include more explanation of the forecast profiles in Fig. 7 b, d.

Increments in the tropics
I don't understand the argument about increments in the tropics. It is said line 17 of the same page that tests "including the use of a second-order velocity balance" are under way. This implies that this balance is not applied in the present system, and therefore that the velocity increments should be zero in the tropical regions. If this is right, how could velocity increments indicate anything as said line 9-10?
The balance in the present system is geostrophic and so velocity increments will be effectively zero near the equator.
We are only talking about tracer increments here and so in the text I have replaced 'increments' with 'tracer increments' to make this clearer.

Figure font sizes
Reviewer #1 also commented on the relatively small font size used for the figures and so we acknowledge that this will need to be increased.

The font sizes used in the figures are either 18 or 20pt but it is the figure scaling employed by the typesetting that causes them to be smaller than the journal text. Ideally these should align with the font-size used for the figure caption.
When the final typesetting is done we shall make sure, in conjunction with the journal, that the fonts used in the figures are clear and in keeping with GMD guidelines.

We have not done so as part of this response because the figure scaling used for this (single column) GMDD discussion document will most likely be different from that used in the (dual column) final version - meaning that any changes we made now may very well need to be redone at the typesetting stage.

The following typos have also been corrected:
p6239, line 5: typo "salinity"
p6254, line 11: typo "will be upgraded"

p6255, line 24: typo “ocean”