Ocean Sci. Discuss., 4, S29–S36, 2007
www.ocean-sci-discuss.net/4/S29/2007/

Ocean Science
Discussions

Interactive comment on “Forecast and analysis assessment through skill scores” *by* M. Tonani et al.

Anonymous Referee #3

Received and published: 13 March 2007

This paper describes the error assessment of the MFSTEP ten-day forecasts through RMS and a skill score based on RMS. The model accuracy is assessed at the whole basin and at various sub-basins, and different depths and times are examined. The model appears to give good results, but I am concerned about the method chosen to evaluate it. I summarize my three major concerns here, and follow with a list of minor comments and a list of more technical comments.

Major comments:

1) The model error is calculated against the model analysis, and not against a set of independent (not assimilated) observations. As a result, unrealistic high SSP (skill score percentage) values are obtained at some zones, as noticed by the authors, because

Full Screen / Esc

Printer-friendly Version

Interactive Discussion

Discussion Paper

in a zone with no observations to be assimilated, the forecast will be very similar to the analysis (their “reality”), but that will not mean that the forecast is good (it is even the contrary, a zone with no observations to be assimilated will have a bad analysis). I would define a good error measure one that gives an absolute estimation of the model error. Otherwise, how can one establish when the skill of the model is high because of high model performance, and when is it high when there is a lack of observations? Ideally, a few observations should be taken apart (i.e. not assimilated into the model) to use them in an independent model error assessment. The authors could weight the benefit of having the highest possible quantity of observations for the assimilation step versus having a more robust error measure.

If not assimilating some of the observations is not a possibility, and if the error assessment is done in hindcast mode, the observations “not yet assimilated” can be used to assess the model error. These observations will not be fully independent from the model, but at least a skill score of “1” would always mean a forecast that reproduces exactly the observations.

2) All error measures used in this work are based on the RMS. The use of different errors measures, such as bias and correlation, would give the authors a more complete view of the model accuracy. As this is, as the authors say, “a first comprehensive evaluation of the quality of the ten-day forecasts produced by the MFS”, then the analysis of the model error should take into account the different types of errors that can appear in a model.

3) A skill score, as described in Murphy (1988) compares the accuracy of the forecast to the accuracy of a reference forecast (assuming that the accuracy of a perfect forecast is equal to zero). The authors choose the persistence as the reference system, but the way they define it does not correspond to the definition in Murphy (1988). A persistence forecast measures the variability of a system with time. What the authors make is to compare the variability of the forecast respect to the analysis at time t=1 (Xf(t) - Xa(t=1), following the notation on the paper). However, to follow Murphy’s definition, the authors

should have examined the variability of the “reality” (the analysis in this paper) with time, i.e, Xa(t=1)-Xa(t). Otherwise, their definition of skill score is not the one given in Murphy (1988), as they cite.

Minor comments:

1) The title is too vague: as it is now we can understand that it is an article about the use of skill scores to assess a model performance, but it is much specific than that (you could mention, for example, model, implementation, zone...)

2) The skill score described in Murphy (1988) is based on the MSE, and not on the RMSE as in this work. Also, the percentage is only a result of multiplying the skill score by 100, so any skill score can be expressed as a percentage. Therefore, I think the name “Percentage Skill Score” is not very informative. Also, in the abstract the authors make reference to Murphy (1993) (references should be avoided in the abstract) when referring to this skill score, and not to Murphy (1988).

3) In the introduction it is stated that the authors “study also the variability of the forecast accuracy due to the seasons”. However, results over less than a year (August 2005 to January 2006) are used in this study, so no complete season (other than fall) can be accurately studied.

4) p. 190; Abstract, l. 10: “The main skill score is computed as the root mean square of the difference between forecast and analysis (FA) and forecast and persistence (FP)” A first reading of this sentence makes the reader think that a skill score computed as the rms between FA and FP is used in the paper. However, after reading section 3 one can see that what is really analyzed in this work is the rms of FA, the rms of FP, and a skill score based on these two rms. Also, I would not call the rms a skill score, but rather an error measure. Maybe the authors can rephrase the sentence to avoid confusion.

5) p. 190, Abstract, l.16: “The rms of FA is always better than FP and the FP rms error

is double than the rms of FA.” This sentence is confusing and a bit redundant, please rephrase.

6) p. 191, l. 9. The authors argue that, while a wide range of skill scores exist, they choose to perform only an rms error assessment to assess the accuracy of the model, and cite two sources (Murphy, 1988 and Demirov et al, 2003) to justify their choice. However, as I said before, I do not think a unique error measure is adequate to evaluate the model presented here, and the works cited do use other error measures apart from the rms. Also, Murphy et al (1988) should be Murphy (1988).

7) p. 192, l.10; “This is a major differences respect the system used in the previous operational system described by Demirov et al., (2003).” Why is this new approach used now (what are the advantages), and what is the difference with Demirov et al (2003)?

8) p. 193, l. 17: “The 10 days of forecast are compared with the corresponding analyses which have been produced two weeks after the forecast production day. All the computations in this study are done using the best available analyses for each considered day”. Place this paragraph sooner in the section.

9) Eqs. (1) and (2): these should read FA(t) = ... and FP(t) = ... and not rms(t) as it is written now, to be consistent with the text.

10) p.194, l. 11; The authors talk about the “persistence skill score” but it should be rather say the “persistence rms”.

11) Figure 2; The data assimilated in the study period is presented. An image showing what is typically assimilated in one cycle would be useful. How does the different coverage in time affect the model performance? (the authors discuss the zones where no data at all is present, but a discussion on how the changes of data coverage in time affect model skill would be interesting)

12) p. 194, l. 18-21: “indicating that the rms of the misfit decays with time at all levels,

showing a beneficial impact of the data on the model accuracy. Thus the analyses are a reasonable good estimate of the reality and we will use them to evaluate the forecast performance." I am not yet convinced by this. In figure 2 we can see that the model misfit decreases with time, and it appears that a steady state is reached at the end of the simulation (this would be more clearly seen in a longer simulation). The analysis can be considered a good estimate of the reality if we are in this steady state (when the errors are small), but otherwise, the model analysis is still adjusting to the observations and therefore it may not be a good estimate of the reality.

Developing more on the previous item, as the initialization of the model seems to affect the model at the beginning of the simulation, shouldn't the error assessment be done after a reasonable period of time to avoid the effect of the initialization on the model skill?

13) p. 196, l. 6-7: "This explains why the rms of FA does not saturate in time but follows the main source of errors that is connected to the atmospheric forcing inaccuracies." I agree that the errors at the surface are mainly due to errors in the atmospheric fields. However, in figure 3 we can see that at 30 and 150m depth the analysis does not saturate either. What happens at these depths, where the effect of the atmospheric fields is not felt within the 10 days of forecast?

14) What is the mean depth of the thermocline in the Mediterranean Sea during the studied period? By looking at figure 5 we can see that the skill score of the temperature at 150 meters is negative at the beginning, and always worse than the temperature skill at 300m. Is the thermocline in the model playing a role in this low skill? A discussion on the skill at this depth on the various zones shown in figure 6 might help understand which zones have a correct water column structure in the model.

15) No discussion of the quantity of data available at depth is done, but I suppose it is much smaller than at the surface. This fact may also affect the skill of the model at depth. By looking at figure 7, we can say that the Gulf of Lions has a worse skill

than the Algerian basin (mostly at depth), and one difference between these two zones is the presence of more in situ data (XBT and ARGO, fig. 2) in the Gulf of Lions. As the authors point out earlier in the article, data assimilation might be inducing an adjustment in the model, but it would be interesting to know how much data goes into each of these zones. Again, a direct comparison with observations is a more efficient way to determine the model error.

16) p. 198, l. 1-2. “This means that the impact of the data assimilation in this region determines an adjustment time longer than in the overall basin.” Discuss why this region would have longer adjustment times.

17) p. 199, l.25-28; “Even if the period of study is not sufficiently long to properly study the variability of the forecast accuracy due to the seasons we believe this is a robust result and it will characterise the forecast errors in all the strongly seasonal basins at mid-latitudes.” This is a strong claim, which is not supported by the data presented here. The model has a good skill in fall, but that doesn’t mean it would be the same in winter and spring. To reach such a conclusion a whole year, at least, should be analyzed.

18) p. 200, l. 2; “it defines a forecast evaluation protocol” As I said before, I don’t think an error assessment solely based on rms is appropriate. I would suggest diversifying the error measures used if this is intended to be used as a routine error assessment.

Technical details:

1) A careful correction of the English is necessary. I do not include a list of errors, but as general advice, unify style (such as, e.g, ten-day forecasts)

2) p. 191, l. 27: Introduce ECMWF

3) p. 192, l.3; “Dobricic et al, 2006 and this volume”: in the references list there is a Dobricic et al (2005) and a Dobricic et al (2006) (this volume)

4) p. 192, l. 13; Add a reference for FGAT

5) p. 192, l. 17; Introduce the acronym “SST” here.

6) p.193, l. 13: sst, sla, xbt and argo: use CAPS

7) p. 194, l. 7; The time “t”, is the day “J” mentioned in the previous section?

8) Table 1: there is a reference to Table 2 in the caption, but there is no table 2 in this paper.

9) Figure 1; Mention what J means in this figure

10) Figure 2; What are the different symbols used in the misfit curves?

11) p. 195, l. 10; Reference to figure 4 should be a reference to figure 3. Conversely, at line 24, reference to figure 3 should be figure 4. Same for page 196, line 6.

12) Figure 4; Explain what the acronyms in the legend mean, and change “std” by standard deviation in the caption.

13) p. 197. l. 16-18. This paragraph is a repetition from what has just been said in the previous lines.

14) In general, figures are too small to be clearly seen at the paper size.

15) The Appendix A is difficult to read: the punctuation is often missing or wrongly placed, and some terms are not explained. For example:

-p. 202. l. 6; “Elimination of undef values” What are these?

-p. 202. l. 9; “Rejection of the whole profile if the distance T and S in the first 150m is greater than 40 m” What is “ distance T and S”?

-What does it mean a flag value of 1? What are the other possible flag values? For example, if it is just 1 (pass) and 0 (fail) then you could simply say that you accept data that pass the quality checks.

-How the “mapped AVHRR SST acquired at CMS” differ from the “AVHRR SST ac-

Full Screen / Esc

Printer-friendly Version

Interactive Discussion

Discussion Paper

quired at GOS-CNR-ISAC”? (p.201, l.10-11)

16) References Pinardi et al (2003) and Demirov et al (2003) have swapped pages (I found these because I looked one of those references, but the authors should check the whole list of references for errors)

Interactive comment on Ocean Sci. Discuss., 4, 189, 2007.