Anonymous Referee #2

We thank the referee#2 for this thorough review.

The paper of P. Yiou, M. Boichu, R. Vautard, M. Vrac, S. Jourdain, E. Garnier, F. Fluteau, and L. Menut presents a new application of an analogue reconstruction methodology for atmospheric fields to the period surrounding an interesting 18th century volcanic eruption. The case period is important, as understanding the effect of aerosols on past climate is an essential step to predicting their effects on climate into the future.

The finding that the Laki volcano did not have a substantial contribution to the subsequent European temperatures is an important one. However, the present study has not made the case that the findings from this reconstruction support this conclusion.

They have not demonstrated that our knowledge of the temperatures during this period is accurate enough. Nor have they shown that transport of radiative-actively aerosol could not have occurred given their reconstructed flows.

I. Substantial comments:

A. Significantly more validation of the method for the reconstructed fields shown in this paper needs to be performed. What is the expected skill for monthly-averaged temperature, SLP, and wind as a function of location when very sparse targets like the Kington (1988) grid are used?

We will make a test on NCEP reanalysis data and removing grid points, in order to check the expected skill from a grid that is similar to the one of Kington (1988), for instance for the year 2010. This will appear in a supplementary material in order to avoid diverting too much the sequence of the paper.

B. More clarity is needed in several areas, particularly the introduction and some figures. These issues are discussed in detail below.

See replies below.

C. Without some physical aerosol modeling, I do not think that the authors' strongly worded finding has been shown: "Hence we find unlikely that the the cold winter in 1783/1784 was caused by the Laki volcanic eruption, because the "memory" of the atmospheric circulation ranges between 5 days to a couple of weeks". At most from the presented evidence, the authors can suggest this hypothesis. The expected aerosol loading from the shorter, less explosive recent 2010 eruption would not be expected to be the same as the long-lasting Laki eruption. A monthly or seasonal mean anomaly arising from radiative forcing can be expressed in a variety of sample paths from the weather noise that we see on a day by day basis. The frequency with which the 2010 analogues are picked (perhaps much higher than chance) in Fig. 3 speaks against the conclusion. Monthly flow fields at 1000 hPa and 500 hPa are not enough to discern the whole atmosphere's transport. A trajectory analysis or diagnosed fluxes throughout the atmosphere would be needed.

The paper essentially illustrates what happens on temperature, and discusses issues that were raised by historians on the summer heatwave and cold winter in France that followed the Laki eruption. The 1000hPa wind reconstruction is a first order illustration of what can be

done. We are aware that a chemistry transport model needs more than surface wind speed and temperature. The incriminated sentence will be rephrased.

A. Specific Validation issues:

1. The Yiou et al. 2013 methodology paper did not show validation for monthly or seasonal averages. The current study depends on the quality of such averages. The observing system experiments of the type presented in Yiou et al. 2013 are needed with validation that is relevant to the present findings.

That paper (in Climate Dynamics) does not show validation for monthly averages, indeed. Such validation appears in Vautard and Yiou (2009). As stated above, we will show the expected skill on NCEP reanalysis data, using the gridpoints of Kington (1988).

2. The Kington (1988) grid is not an averaged grid. The values represent grid points. The NCEP-NCAR reanalysis data are also grid point values. Why were the NCEP data averaged instead of taken at the same grid for the analogue target to base RMS measure?

As explained to the first reviewer, the grid values stem from meteorological stations averaged on 5x5 cells. Averaging the NCEP SLP over 5x5 cells seems the most reasonable approach, rather than the nearest neighbor.

3. While sparse, there is substantially more observational data to use for validation and comparison. For a paper making these very interesting claims, one should maximize the use of the available observations. This would help to assess the results.

Significant collections of marine observations that can be used for independent crossvalidation of the wind and temperature are in the ICOADS Release 2.5, which includes the Climatology of the World Oceans (CLIWOC, Garcia-Herrera et al, Climatic Change, 2005) dataset spanning 1750-1850. These wind, temperature, and pressure data are available from NCAR at http://rda.ucar.edu/datasets/ds540.0/. As the reconstruction is daily, the original marine observations and their monthly averages can be examined for the comparison with the reconstructed variables. Monthly temperature reconstructions from the Berkeley Earth Surface Temperature dataset (1753-2011) are also available (http://berkeleyearth.org/)

We took the time to download the ICOADS data for the north Atlantic, and study its content. The ICOADS dataset is not that easy to use for comparisons, because it is based on ship information, that move at a speed that is proportional to wind. Hence, making monthly averages requires statistical procedures and choices that are potentially debatable.

We also looked at the Berkeley Earth Surface Temperature dataset (the dataset actually contains anomalies of temperatures). The time variations are indeed similar to those of meteorological time series.

4. Is there any sensitivity to the base dataset that is used? ERA 40, 20CR, and the newly available JRA-55 all span a significant period of the base NCEP-NCAR reanalysis dataset. Since the calculation is so much less computationally expensive than data assimilation, showing the sensitivity to the base dataset seems like a reasonable part of the investigation into the reliability and robustness of the results.

We made the comparison by replacing NCEP by 20CR. The reconstructions are very similar, although the analogues can be picked in a much longer longer database. This will be mentioned in the text, and we will provide a supplementary figure.

5. The authors need to show or discuss error bars or uncertainty in all cases where possible (e.g., Fig. 4, 5, 6, 7, 8).

The standard deviations across the analogue fields will be mentioned.

6. The authors also need to estimate whether the validation data where available are consistent with the internally estimated errors.

This will be done by comparing grid point reconstructions with individual validation data.

7. The authors need to present maps corresponding to Fig. 4 of the standard deviation or some other measure of spread of the reconstruction. How big are the anomalies compared to an internally estimated uncertainty? This is indicated to some extent in Fig. 5, but I am confused on the interpretation of this figure. The grey boxes make it appear as though at no time are the anomalies significantly different from 0. However, the dashed lines seem very consistent with the median, so perhaps the grey is overemphasizing the uncertainty.

The grey boxes reflect the probability distributions of daily reconstructions for all analogues, within a given month. The ranges can be large. The lines are medians over daily values for each month and each analogue.

8. In addition to showing the regional consistency of the station data in Fig. 6, a more direct comparison of the reconstruction with the stations is needed. Averaging and computing the standard deviation of the 20 analogue reconstructions at the station locations and comparing to the station average would help the reader understand the internally estimated reliability of the reconstruction.

OK, this will be done and a panel will be added to Fig. 6.

9. A direct comparison with a reconstruction of 2009/2010 is also needed. Just how well would this method have detected the cold winter in 2010? Are the anomalies in 1784 large compared to what this method reconstructs?

The reconstruction of 2009/2010 appears in Cattiaux et al. (2010). We will discuss the amplitude of the reconstructed anomalies.

B. Specific Clarity issues:

1. As the methodology of Yiou et al. 2013 is central to this paper, I suggest a more in-depth summary of the methodology, perhaps including the helpful figure of Yiou et al. 2013 Fig. 1. This should help make the paper more readable and self-contained.

This is also suggested by the comment of D. Wheeler and reviewer 1. We will add a methodological appendix.

2. Could you explain in this paper how the overlap is handled? Perhaps I am misunderstanding, but it seems that are actually 2 “first” analogues for many days. One first analogue from the first window and one from the second window that includes a particular day.

This will be recalled in the text.

3. Somewhere the trade-off of the expected loss of skill from not using data assimilation compared to the cost of using it should be mentioned. The results of NOAA's 20th Century Reanalysis (Compo et al., QJRMS, 2011, see, e.g., Fig. 9) and ECMWF's ERA-20C (Poli et al., 2013, http://www.ecmwf.int/publications/library/do/references/list/782009, see, e.g., Fig. 24a) appear to be substantially better than the analogue method is giving in Yiou et al. 2013, but at a significant computational cost. Yiou et al. 2013 did not point out that the analogue method is giving up something like 30% of the daily variance compared to using data assimilation (correlations averaging 0.9 or higher compared to averaging about 0.7 or less for upper-level geopotential height).

There is indeed a necessary trade-off, because the weather is not "recomputed" by assimilating data, but we estimate a weather that has been already computed (in reanalysis) that is compatible with early observations of SLP. This is a heavier constraint than what is done in 20CR, where the model creates its own (self consistent) meteorology.

4. Another trade-off of the Yiou et al. method that should be mentioned is the lack of direct dependence on a dynamical numerical model. This feature should be highlighted, as the confluence of model error and observation density may affect the reliability of some trends in reanalysis (e.g., Ferguson and Villarini, J. Geophys. Res., 2012).

The lack of direct dependence on a dynamical numerical model will be highlighted in the text. The homogeneity of our reconstruction indeed mainly depends on the homogeneity of the Kington dataset, which is debatable.

5. I am confused by the lower panel of Fig. 2. Are the scores related to the monthly average of the Kington dataset or to monthly averages of the daily scores? Regardless, since medians and averages are being shown in subsequent , the score of the median analogue needs to be reported, also.

No. As stated earlier, the grey boxes reflect the probability distributions of daily reconstructions for all analogues, within a given month. The ranges can be large. The lines are medians over daily values for each month and each analogue.

II. Technical comments:

1. pg 5162 line 5. Is surface temperature or 2 m air temperature being used and plotted throughout the paper? They are not the same. Some details of the station measurements are needed. At what height were they, for example? Are they screened in any way? Regardless, as the temperature fields described come from NCEP-NCAR reanalysis, they are generated on an irregular T62 Gaussian grid and not a 2.5 degree grid (Kalnay et al. 1996).

Those details will be provided in the text.

2. pg 5162 line 10. I did not receive access to the Supplementary Movie.

Sorry, we forgot to provide the address:

https://cloud.lsce.ipsl.fr/public.php?service=files&t=cd64552c3de72b5156f7021dbc15f929

3. The meaning of the circle symbols in Fig. 1 is not given in the caption. What do they mean?

See reply to referee#1: The circles are for the outliers. Classically, the upper (lower) whiskers of the boxplot are the minimum (maximum) between 1.5 times the interquartile range and the maximum (minimum) value. The circles indicate that there are skewed distribution tails.

4. pg 5161 line 3 and throughout. The use of the term "target" may be confusing to some readers. I think the authors mean "predictor".

Referee#1 made the same remark. The terminology will be changed.

5. The arrows in Figs. 7 and 8 are not readable in some maps. I suggest changing the reference arrow size to one that allows the reader to see the flow in the most important panels being emphasized in the text.

The size of the arrows will be reduced to enhance readability.

6. The introductory paragraph laying out the paper is confusing in its current order. Please mention sections in the same numerical order as presented.

OK, we will pay more attention to this.

7. pg 5163. I think you mean "The fields corresponding to the dates of the 20 analogues allow us to make three dimensional reconstructions of the atmospheric flow and temperature for that period." The dates themselves do not have information about the flow or temperature.

OK, right.

8. pg 5163. Why are the Kington data being used to estimate the autocorrelation structure of SLP? Much more accurate data are available for the modern period from the new reanalysis datasets, such as ERA-Interim, for a longer time period than Kington. I expect that the error bars in Fig. 1 are large because the sample size is small, and the Kington data contain errors. A figure based on modern data should also be shown.

Since the reconstruction is based on the Kington dataset, we feel that it is more appropriate to compute the auto-correlation on this dataset. This also implies that the reconstructions yield auto-correlation properties that are similar to the one of Kington. We will perform the same exercise with NCEP data.

9. pg. 5164. The paragraph beginning o line 24 is not comprehensible.

OK. The sentences will be shortened in order to convey only one message.

10. pg 5165 line 27. In what way are the maps shown in Fig. 4 "spatially averaged"?

Sorry, this had escaped our scrutiny. The text will be corrected to reflect the figure caption.