Review of the manuscript:

Bimodality in Ensemble Forecasts of 2-Meter Temperature: Event Aggregation

by Bertossa et al.

This manuscript addresses an interesting topic. Based on the comments below, I recommend publication subject to major revisions.

Major comments:

1 The methodology used in this study is perhaps unnecessarily complicated and uses lots of arbitrary thresholds and other subjective elements. This undermines the robustness of the study. This is beyond something that is not elegant – it is cumbersome and hard to defend. I recommend that the Authors scrutinize their methodology and eliminate unnecessary conditions or base their choices on more objective conditions. Examples include: l.111, 116-117, 128, 136-137, 154 (subjective choice of 3 regions), 160, 166, 179-185. Some of the arbitrary choices are necessitated by earlier subjective conditions etc. Several of these choices appear to be made so a desired outcome is achieved (e.g., l.160, 179-185), undermining the generality of some of the findings. For further details, see detailed comments below.

2 The study presents results suggesting bimodality in the distribution of ensemble members. This condition, however, is never established objectively. I recommend that the Authors consider developing an algorithm to test whether the results they present are statistically (in)distinguishable from surrogate data generated using a null hypothesis that ensemble forecasts are normally distributed. Are results over the entire sample (i.e., over selectively chosen data) indistinguishable from randomly generated samples from a normal distribution? Without such a test, the results are not convincing and would be appropriate to present only as a short note.

3 In my view, no major conclusion appears to emerge from Section 3. The discussion is mostly descriptive, with no insight into the dynamics leading to the emergence of bimodality. As the results in Section 3 may be limited in significance, I find the number of figures (10) excessive. I recommend that the Authors significantly reduce the number of figures and resort to using short verbal statements to describe less important results. Also, as some of the results for the three regions studied are similar, results for the three regions could be compared and discussed together, shortening the text.

4 The Discussion in Section 4 is interesting but appears somewhat preliminary. A more logical organization of the material and a shorter presentation would improve this section.

5 Apart from its last paragraph, material in the Conclusions section also appears to be disorganized and premature. As the manuscript mostly presents hypotheses without objective validation, unless more evidence is added, some of the statements (e.g., l.420-421 should be tempered accordingly. Depending on how the Authors respond to the reviews, this section also needs a major revision.

6 The quality of the presentation is mixed. There are some parts in the Introduction that are well written. The bulk of the manuscript, however, needs significant improvements before possible publication. See detailed comments below for some examples. Some suggestions:
a) Clearly define all terms used in the study. Examples include: size, occupancy, points, forecast in the description of the methodology. Some terms may be poorly named, using commonly used words with a different meaning. In such a case, clear definitions are especially important.
b) Figures should be defined in their captions but the discussion of the results should appear in the text.
c) Replace hypothetical examples used in some figures (e.g., Fig 2) with examples from the actual data.

Detailed comments:

1) L. 25-26: “since the growth of perturbation variance in an ensemble”
2) Sentence in l. 29-30: this may need some additional explanation and/or references
3) L.34 – not clear
4) L.35 – “dressing methods” – please explain (or provide proper references)
5) L.42: “guide our analysis: (1) Is bimodality…”
6) L.47 – not clear
7) L.48, l.67, “ensemble spread” is customarily used to refer to the square root of the variance of ensemble members around their mean. I suggest you use “the distribution of ensemble members” or “ensemble distribution” when referring to specific behavior of such distributions, like bimodality
8) L.49: “answer question (1)”
9) L.51, “contribute to answering question (2)”
10) L.52. “naturally addresses question (3)”
11) L.53: Clusters need to be defined before they are discussed any further

12) L.65 – “concludes the study”
13) L.75 - how “groups of forecasts” differ from clusters?
14) L. 78-79 – Why is it so?
15) L.80 - Please clarify “based on their 2m temperature values”
16) Section 2.1 – The text is hard to follow, which prevents the reader to assess the material presented here.
17) L.92, end of sentence – not clear
18) L.111 – “few dozen forecasts” – why is this a problem?
19) L.128 - is this redundant, given a similar earlier statement?
20) L.140-141: not clear why this may be the case? This may be suggestive of a causal link, but may not necessarily be an indicator that full life cycles are assessed?
21) L.148 – Fig. 3?
22) L.150, “developing around true dynamical events” - an alternative explanation that needs to be evaluated by statistical tests is whether bimodality may emerge in normally distributed 2m temperature data due to sampling fluctuations, see major comment 2.
23) L.175-178: Admittedly, the results indicate spatiotemporal consistency. Due to dynamically conditioned covariance in meteorological data, this, however, would also appear if bimodality is a result of pure statistical sampling fluctuations in normally distributed data (“normal distribution hypothesis”). Can you substantiate your claim of dynamical origin for the emergence of bimodal signals by refuting a normal distribution hypothesis via specially designed tests (see major comment 2)?
24) L.198-199: perhaps not “optimizing” by refining or adding arbitrary choices, but rather revising the methodology may be needed to reach more firm conclusions?
25) L.204 – “lead to the emergence of bimodality”
26) L.238 – “eastward propagation”
27) L.240, Fig. 7 and similar figures: this is a rather unusual presentation of the data. The 2 panels use different formats, making a straight comparison difficult. Would it be more informative to show anomalous conditions (from the mean of the ensemble) in the same format for both modes?
28) L.242, “medoid” – not clear what this refers to
29) L.269, “exhibiting usefulness” - Admittedly, one of the two modes is found to be associated with a synoptic development documented earlier. In what sense does this make the methodology useful?
30) L.301-310: these are two interesting hypotheses, just as topographic forcing and Rossby-wave breaking ideas introduced later. The four hypotheses presented appear reasonable, and beg for validation. I wonder if the Authors

could experimentally confirm any of these hypotheses by examining sea ice and other relevant data? The discussion in this subsection is relatively long and most of it is speculative. Without such an analysis these are not more than interesting suggestions. I would like to suggest that the authors shorten this subsection, unless they wish to present evidence confirming or denying each hypothesis presented. As an example, l.369-370 could be removed without loss of general information.

31) L.354, " Reflective of what is depicted in Fig. 12" – superfluous, consider deleting
32) L.357, "in the South American region"
33) L.361, "The typical speed of the bimodal events as indicated by Fig. 12 is roughly 5-10 ms−1," - redundant, given similar statement above; consider consolidating the two statements about speed.
34) L.382-383 – interesting hypotheses mentioned here, though without validation, they have limited value.
35) L.391, "model spread may be adjusted appropriately to improve forecasts" – please clarify
36) L.394, "metastable" – is this a good choice of word here?
37) L.410-411: "recurrence and quasi-stationarity"
38) L.422-426 - I find this confusing. Bimodality is a characteristic of forecasts, not weather events, isn't that true? Please reconsider.
39) L.428 - Is this contradictory with the discussion of multiple causes of bimodal forecasts in Section 3?
40) L.440: Has this been presented in Section 3? It appears to belong there.
41) L.445: "as an area of frequent Rossby…"
42) L.447-448 - this sounds a bit far-fetched and not supported by this study. Consider removing or qualifying this statement.
43) L.449 - Yes, more insight into this basic question would certainly make the manuscript more substantive.