**Major revision of JCLI-D-150219**

Comments by reviewer #2

Overall comments

1. The authors claim to assess the mean climatology of California "including temperature and precipitation", but in fact only temperature and precipitation are analysed.

X. Huang: Did not get this.

2. The general presentation of the paper is chaotic, and despite the fact that a lot of analysis has been done, the authors fail to find a clear and simple way to achieve their objectives.

Major comments

1. The authors are not clear about their understanding of "dynamical downscaling strategy". There are two ways to obtain an RCM: by using a limited area model (LAM) or by using a variable resolution GCM (VRGCM) (Laprise, 2008, Regional Climate Modelling, J. of Comp. Phys). This issue needs to be clarified and the text needs to be corrected accordingly.

X. Huang: Agree and it has been corrected. Further check the paper at <http://www.sciencedirect.com/science/article/pii/S0021999106005407>

2. Within VRGCM there are also different strategies of achieving high-resolution over the area of interest - stretched-grid models or grid refinement (as mentioned in the paper by Zarzycki et al (2015), which was cited in the manuscript). Which of these strategies are used in CESM?

X. Huang: It is grid refinement. This note has been added.

3. The paper is poorly organized. In the "Introduction" part, the authors speak about CESM, continue with their objective but forget about the comparison with a uniform resolution model (mentioned in abstract); the authors then continue with the second model and finish with the importance of studying California's climate. They should present the importance of studying California's climate at the beginning, which would make their objective more understandable. The authors need only to tell us which models they are going to use, and they should leave the detailed description of these models to be addressed in section 2.

X. Huang: I am not sure about the latter part of this comment as marked in red color. Anyway, if moving this part at the beginning, I will add more details about this aspect then. Correction has been made followed by the remaining parts of this comment.

4. Section 2 should also be reorganised. First, they must present varres-CESM; this is a fully coupled system with separate models for atmosphere, ocean, land and sea ice. How is this coupled model working? What are the variables that different modules exchange? Besides the affirmation (put in the wrong place - the introduction of section 2.a instead of 2.a.2) that the ocean model is disabled and the model is forced with prescribed SST and ice concentration, it is not clear which module of this system has been run or not.

Secondly, they should continue with WRF (a LAM model); here it is not clear how "two-way nesting" is enabled when the simulation at 9 km is performed.

In section 2a the authors must also include the uniform resolution CESM, which is a model not only a dataset. No details are provided about this, only a reference with the specification that some parameters in the paper are different from "public release". Which parameters? Is this a negative influence on your experiments?

X. Huang: These two parts have been reorganized. Further changes need to be made.

5. Section 2 is called "Models and Methodology" but no methodology is presented. There is only a brief mention of post processing but it is not clear what the authors are going to analyse.

X. Huang: This part will be rewritten.

6. Why is the topography included in the section "Results"? The topography is a forcing that should be included in the previous section. They should also give more details about the configuration used by each model. Something appears on page 8 about CESM but nothing is said about WRF. Figure 3 should be presented there.

X. Huang: Added.

7.  The results are just presented but they are not discussed. The authors frequently speak about "bias" without mention to which dataset they refer. They computed scores against three datasets as shown in Table 2 and 3, however the authors present the average fields for models and observational datasets without explaining why these are presented. Also, bias is presented only for two datasets, but it is not clear why only these two datasets were selected. The authors begin by presenting Tmax and Tmin for summer, they then continue with Tavg for summer and winter and finish with Tmax and Tmin for winter. This way of presenting the results can prove to be confusing for readers. For example, what was the purpose of presenting the NARR dataset? No biases against this dataset are shown and only three times the word NARR is mentioned (lines 226, 228 and 250).  NARR is mentioned only to say that comparisons are not shown or that there is "bias in NARR", but without mention to what

the bias is related. Also, in the figures representing the seasonal cycle there is not a quantification of the standard deviation bars, the readers have to compare by themselves in order to see which one is important or not.

X. Huang: I will add more interpretation of the results and reorganize them to be readable. Also, the standard deviation bars will be added in a new table or with quantitative words.

8. When comparing variable resolution configuration with uniform resolution configuration, the authors refer again to "bias" but there isn't any bias that is presented and the authors do not say to which observations they are referring to see the magnitude of the bias. They compare the small-scale features for temperature but have they compared the large-scale futures of these two configurations? If this comparison is not shown, reference to it should at least be made.

X. Huang: I will add the referring observation and explicitly point to the figure displaying the obs. showed before. Large-scale features of these configurations will be referred to {zarzycki2014multidecadal, zarzycki2015effects} and states some important conclusions of this aspect. Is it necessary to add another part of large-scale features comparison?

9. Finally, there are many acronyms that are used in the text, but no explanation as to what they stand for is given. Some of the acronyms used are: WRF-ARF, FAMIPCS5, WSM.

X. Huang: Checked.

Comments by reviewer #3

1. The lack of consistent definitions and abundant use of vague expressions is more obvious.

**Minor points:**

**1.** One similar diagram for precipitation should also be retained as Fig. 4. Verbal descriptions of these pictures take up a considerable part of the paper. These are not useful. It may be necessary to do visual model evaluation extensively and even for figures not shown in the paper. It is not clear why only some time intervals are investigated and why there is a preference for 3 month averaging. It is sufficient to describe only examples. If a general report on visual inspection is considered necessary, the reader should be asked to take it on thrust and this would be one or two sentences. The examples for visual inspection should not take up more than 1 page. It is of no interest to fill many pages with visual descriptions. The authors can point to an internet page giving many output results for readers who want to go through them. If climate investigations relay on this, they will be quite subjective. In these figures everybody finds something he likes about the models and something he likes not. The deleted information should be replaced by more statistical information of the kind given in the tables and figs 9 and 12.

X. Huang: I will correct these.

2. A good point is the use of rms and variance information. While for rms it may be just sufficient to define the area, the whole statistical evaluation procedure needs to be defined precisely. There need to be a section with formula to do this.

X. Huang: Ok, I will add a section in the methodology part to give the formula.

The authors and people in their institute may be in a certain habit to do things. This is, however, not yet generally accepted. For example using a “climatology” based on 30 rather than 26 years is an acceptable alternative.

X. Huang: Does this really matter?

For example at the end of page 11 they speak of Tmax climatology, when the reader supposes that just the average is meant. I suggest that the simple language is used and all words used are defined. There is more than one way to form an average. For the variance they make believe that this is an unique thing while in reality this needs to be defined. The statistical model (s) needs to be defined precisely and the confidence interval stated. What impact on the result does the latter have? The ensemble to take sigma needs to be defined. For example we deal with parameters which are defined by a statistical model and a possibly random component. The word bias should be defined. From the tables It appears to me that the average fields over the whole model area and integrqtion time is meant. But then the word regional bias is used. When discussing moisture, bias is used as a synonym for error.

X. Huang: Ok, I will define all the statistical terms I used.

The model the authors prefer is a dependence on time and space. If the whole original ensemble is used, then the variance is overestimated, which as such is not necessarily bad.

X. Huang: Did not get this.

More options than currently in the paper should be given. For example maps of variance could be presented. The statistical significance of different statements is really interesting, but never given.

X. Huang: Ok, I will add these.

The choice of different time intervals, other than the 3 month or one month should be discussed. The 26 jears integrated are probably dictated by the resources available. If the authors think that this just happens to be the right ensemble to investigate climate they need to prove this statistically. What is the error margin for 10 years? What is the one year fluctuation? This means how big needs an effect to be to make this year exceptional. A trend is possible in the data of 26 years. The authors need to discuss possibilities to include this in statistical modelling.

X. Huang: Ok, I will do this time analysis. Did not get the sentence marked in red.

3. With respect to 2 I would like to clarify: If a paper is about an new model, such as varres and provides just one case showing that the model is reasonable, this would be acceptable and the small sample would not stand in the way of publication. Here, however, we have a large sample allowing refined statistical investigations. Therefore the statistical tools need to be sharpened to the degree that future investigators of related climate questions, depending on the accuracy required, can decide if they need 5, 10 or 25 years of integration. Therefore a variety of statistics need to be compared with respect to the accuracy achieved.

X. Huang: Continued from the above time analysis. I will try to analyse the time length requirement with respect to the statistical accuracy.

4. While the models used are discussed using sufficient references, too few papers are cited dealing with climate evaluation.

X. Huang: Ok, I will try to add more citation in this aspect.

5. In page 8 pp the authors discuss the datasets used for verification, which is rather necessary. Unfortunately this comparison remains rather qualitative. Only in Table 2 it appears indirectly that these differ among themselves more than desirable. If this is just an effect of different resolutions remains unclear. It is necessary to apply statistical evaluation methods, such as variance information for each dataset and of their difference. The whole statistical model should account for the fact that the verificatioin data sets have errors.

X. Huang: Ok, I will add this. Not sure what the statistical model is needed to account for the fact that the verificatioin data sets have errors so far.

6. From the summary and introduction I learn that the authors specifically want to investigate the variable resolution aspects, and above all the new 1:2 refinement now possible with the SE version of the HOMME model. In this respect not even a discussion is presented. The 2 way interaction of WRF is not conserving, while the 1:2 refinement is conserving. This needs to be discussed and diagnostics near the boundary with respect to conservation done. In particular a comparison of point forecasts and climate averaging needs to be done, to investigate the deviation of the non conserving model and if the climate averaging takes this effect away. Many people think that 1:2 refinement is not possible or not reasonable. Therefore the model behaviour needs to be investigated to show if in point forecasts or climate averaged results strange features coming from sudden model refinement are present. How do fields near the refinement boundary look, as compared to uniform resolution? I myself have always considered 1:2 as possible, but many colleagues don’t and go to great lengths to create models which change resolution smoothly. For the benefit of this discussion results should be given, showing the impact of 1:2 resolution refinement as compared to uniform resolution. The differences in model performance seen and described in the paper are very likely caused by other model differences than the specific refinement procedure used in the models.

X. Huang: Good point. Need to further discuss this.

7. I know that the grids can be seen by going to the cited papers. However I suggest to make it easy for the reader and plot a section of the grid with 1:2 refinement.

X. Huang: Ok.

8.Rather vague words are frequently used: ‘the error is relatively acceptable overall”, ”no significant differences can be found”’ ,” uncertainty …is unlikely impacting our results”, “some mismatch occurs between the observational datasets” and many, many more. These statements must be replaced by proper statistical evaluation, such as: The error is below… and not significant using statistical model…, The differences remain under the significance level…., The mismatch between datasets is below the measure … and this does not destroy the error levels of … found. This will depend of course on what the authors found, which I could not quite make out. Very likely the situation is more complicated than these simple statements make believe and needs to be described statistically precise. Sometimes the authors find differences, for example with WRF 9 km. In such cases the reader wants to know if one of the models is in error or if the difference is within the range of ensemble variation. In other words: are these differences statistically significant. I find it remarkable that the high resolution observational data set also sticks out compared to the others.

X. Huang: I will correct these vague words.

9. Fig 9 is useful as it makes a little step towards the requested statistical evaluation. I could not make out, what the confidence intervals really mean. Why are there so many and different for the different models and observational systems. Some models are clearly outside the confidence intervals of observations. I could not find out enough about the definitions used to be able to derive the physical interpretation of this.

X. Huang: I will explain this figure in more details.

10. As my own interest is improving the precipitation simulations of models, I expect substantial errors in simulation, but still was surprised about the amount of errors and about the huge error bars. Or are this margins of variation? I really would like to understand these results better. For the solid coloured error bars the reader can at least hypothesise, but I did not find any clue what the dashed bars mean. In fig. 11 both high resolution observations and models have increased precipitation in places. Is this for real and to what degree? The datasets and simulations need to be statistically discussed under this point of view. Fig. 12 gives huge error bars or bars of variation. Most models are together more than the error bars indicate, in particular for mountainous area. The error bars are much wider than the difference between models. What can we say about statistically relevant differences? Model error bars vary between 1 and 12, observations between 1 and 7. Can we say anything with confidence about precipitation? Can the WRF 9km result statistically be proven to be in error? By casually regarding fig. 12 I was in doubt if climatological evaluation of precipitation by the models used makes sense, even though the models among themselves stay together. Much more explanation is necessary.

X. Huang: I don’t think the deviation value is the error. I really need to state more clearly in words. I did not get why the confidence about precipitation will be a problem.

**Major points**

1. Statistical methods need to be much more precise and must be described in much more detail and this should be in a separate section containing formula.

X. Huang: As in the first part of Minor points No. 2.

2. I would like to make clear, that I do not request over the top rigor in statistical methods. However, the reader should be given the tools and information to decide if a difference observed is real or within the natural variation. It should also be described how errors and significance depend on the sample used. A reader may be interested to know if for his own planned experiments and the accuracy he seeks a 5, 10 or 25 year run is reasonable. He may also want to know to what degree results may become more accurate using an ensemble of climate runs. I do not ask the authors to provide such ensembles. There is enough potential for a good paper using the results they have. However, correct statistics should make it possible to obtain information, how accuracy depends on the size of an experiment.

X. Huang: Ok. As in the last part of Minor points No. 2.

3. Just one out of many, many examples from page 16, bottom: “ …. Varres-CESM and WRF 27 km show satisfactory performance…., although these conclusions are also constrained by observational uncertainty.” What satisfies the authors is not interesting. They should make precise statements and give their statistical significance and the statistical model should take account of observational uncertainties.

X. Huang: Got it. As mentioned in the Minor points No. 8.

4. Let me try to convince the authors of the necessity of statistics by a comparison to numerical forecasts. For these the forcastibiity, depending on forecast element and lead time is well known. This was not always so. I remember times when on the basis of one or two experiments, which the developer liked, new model versions were introduced. ECMWF changed this and made model changes on the basis of a large statistical experimental basis. This made possible to pinpoint relevant model changes and was the basis of ECMWF’s early success. Correct statistical procedures in climate will be able to pinpoint the model changes improving climate performance, the climate errors of existing models and the relevance of parameters for climate change. Variance information is for climate even more important than in forecasting. I remember a visit of the government climate institute of a small country, whose Government wanted to make an enormous investment depending on the question of how fast climate change will arrive. The expectation value of the interesting parameter (sea level height) was clear enough. However, the scientists went to great length to investigate the probability of a development outside the norm. Given the huge investment of their government and that in their opinion the future of their country was at stake, they were right to do this effort. Almost certainly the statement, that their conclusions were “satisfactory”, would not have been bought by this particular government.

X. Huang: Got it.

5.The qualitative and picture describing effort of the authors could also lead to a paper. This would mean that just a sentence like “ visual expectation of many outputs of this and that kind lead to the conclusion that the models compare in this and that way. I would have to recommend such a paper for publication, but it should have no more than 2 pages. I would not like this approach and can imagine that the authors also don’t.

X. Huang: Ok. As mentioned in the Minor points No. 1.

6. If the authors decide to put more effort into statistics, the volume of the paper should not increase, but less meaningful output be replaced by output describing the statistics. Just some steps in the direction of statistical evaluation do not justify publication. I also do not require to arrive at a completete statistical evaluation package for local climate. However, to stay in the picture, a good day of hiking, rather than some steps is required.

X. Huang: Ok.

7.The authors speculate on possible causes for errors. Readers may want to do experiments on model improvements and to investigate the impact of model parameters, such as changes in agriculture. To repeat the full experiments of the authors Is a big effort. Therefore I suggest that the authors relate their findings to sets of model forecasts. If a model run is considered as “truth”, then for example a 11 day forecast may be considered as a ten day forecast with modified initial values. In this way the author’s run contains many forecasts. If the lead time of such forecasts is large enough to loose forecast skill: Is the average of such forecasts related to the long year average? If this would be the case, sensitivity studies can be done by a set of forecasts, rather than one long run.

X. Huang: What is a set of forecasts? Does this mean I need to get data sets of forecasts from other resources?