**Time-Series Prediction Approaches to Forecasting Deformation in Sentinel-1 InSAR Data**

**P. Hill, J. Biggs, V. Ponce-Lopez and D. Bull**

**Response to Reviewers’ Comments**

We would like to thank the associate editor and both the reviewers for their helpful and very detailed comments on our paper. We have addressed all of their comments in detail, updating the manuscript where necessary. We believe that in addressing the issues raised, that the paper is considerably wider in scope and has led to stronger conclusions and outcomes.

The following document repeats the reviewers’ comments with in-depth responses to each issue given in red.

Where possible changes in the manuscript text suggested by each reviewer is highlighted in bold.

**Reviewer #1 (Formal Review for Authors (shown to authors)):**

**Short summary**

The authors present a study in which two problems are described. In the first, they present three methods (sinusoid fitting, STL decomposition, and ACF) that can be used to determine if pixels are exhibiting seasonal variations. They provide an interesting and convincing figure showing the classification of pixels using these methods for an area in the UK. In the second, they explore the ability of methods such as LSTMs, sinusoids, and SARIMA to forecast the behaviour of pixels for up to nine months. Additionally, they use the methods described in first part of the

manuscript to further explore which of the forecasting methods are best suited to pixels or areas that exhibit seasonal variations. They conclude that the problem is challenging, but that several of the machine learning methods outperform other models on timescales of less than three months.

**Major comments**

* The introduction to section 3.1 (107 - 124) seems rather hard to follow in places (though is interesting). Ideally, I would like to see this section rewritten, with particular attention to:
  + 109 and 113: Use of "we start", and then "first" when describing what the authors did.
  + 111: To make clearer, it might be nice to add a reference to section 3.1.1 here (sinusoid fitting).
  + 113: Authors mention that they review variety of methods and summarise in table S1. Where is the actual reviewing done, and not just the summarising?
  + 119: STL - again, add a reference to section 3.1.2 perhaps to aid in clarity.
* **98 - 104: It would perhaps be worthwhile to add a figure (or extend figure 1) to show the time series for at least some of points 1-8, rather than simply describe them in the text.** These time series are shown in figure 8. We have included some text to the caption of figure 1 directing the reader to the actual time series representations.
* Figure 4: I believe this figure does much to explain how the data are organised for the forecasting part of the paper. However, readers may find it hard to follow. For example:
  + I assume Nx = 10 corresponds to months (though in the text it says 9 was used - line 222). The manuscript may be more readable if either days or months were used consistently as the units for time. However, on 193-194, Nx is the "number of past observations"?
  + You only label two examples of y and y^, and these happen to be in the testing data. However, as I understand it, for each sliding window there is also a y and y^, so many of these are during the training stage (and are, infact, what the model is trained on). Perhaps this figure could be redesigned to avoid suggesting to a reader that y and y^ only occur in the testing stage.
  + The x label may be clearer if it were consistent with the y label (e.g. "quantity (units)").
  + Tick label font is perhaps too small (especially on the upper left subplots)
* Multisignal LSTM.: A description of how the data are used in this case may aid a reader to reproduce the experiment (as an aside, could the two types of LSTMs be referred to consistently - either uni/multi variate or uni/multi signal? Different terms are used in 4.1.1 and 4.1.2. ). Specifically:
  + 231: "from a set of training signals" - I interpret this as meaning that instead of using a single training signal, you are now using multiple? If so, it might be clearer to quantify this.
  + 234: As above.
  + 234: "The testing data remains the same" - Perhaps clarify this – I assume, just predicting the time series for one pixel?
  + 234: It may make understanding this easier if the size of the tensor used to train the model is explicitly stated. I assume X would be (None x 10 x 6 ), or perhaps (None x 40 x 6) [note that I can't work out how many data points are used for each pixel, tying in with the previous Nx confusion, and "None" would be the batch size, as per ML terminology]? Y would be (None x 10 x 1), or (None x 40 x 1).

**Minor comments:**

* **49: Might be worth adding Gaddes et al., (2018) reference to this list as it describes detecting the 2015 eruption of Wolf volcano.** Added
* **52-54: For completeness, it would be nice to see references to examples in these fields.** Added
* 76: Figure 17? Updated to correct number
* **96: I don't quite follow this - the text describes 15mm/yr, but the figure scale ranges from -2 to 2?** 15mm/yr is the maximum downward displacement. The red colour on the colour scale on the image shows 2mm/yr or more.
* **Figure 1:**
  + **North arrow.**
  + **It would also be helpful to see where this is. E.g., an inset overview map, or lat/lon axes (which would negate the need for a north arrow).**
  + **Showing the area covered by Figure 2 may also be useful.**

We’ve included an inset overview map as suggested.

* **141: is "First Loess" smoothing a thing, or was Loess smoothing first applied? Could be clearer for a JGR:SE reader who may not have encountered these terms before.** We’ve added a comma between “First” and “Loess” to clarify.
* **144: S is not explicitly defined, whereas L is. It might be worth tying these terms into the description given on 141-142.** The definition of S, T and R have been updated.
* **Figure S1: Tick labels have been cropped off.** Updated
* **Figure 2:**
  + **It may be worth changing the colours used here - the yellow points are very hard to see against light backgrounds (exemplified in the legend).** Each colour needs to show up on both light and dark areas. On examination, all the yellow dots are able to be seen on the image. However, they do not show up well on the legend. The legend has been updated to show the yellow dot index.
  + **It feels a little odd for three measures used in the same manner to be given vastly different amounts of figure space. Also, the area covered appears different between plots.** The ACF plot is given a prominent position and more area on the page as it was chosen in subsequent experiments and the legend was extended to show the confidence intervals of the different coloured bands.
  + **Would it not be logical for the subplots to be ordered in the same manner as the text (a = sin, b = STL, c = ACF).** The order was given for the reason above.
  + **Scale?** Added
* 164: Is there a reason why buildings should have a strong seasonal signal? Note - this is mentioned later in the paper (Discussion, I think), but it may benefit the paper to bring it forwards to here.
* **Figure 3:**
  + **What is the justification of these plots not being square?** Made square
  + **As mentioned before, yellow is hard to see.** The yellow markers have been changed to magenta
* 175: I assume ACF is chosen as it looks best in Figure 2? If so, it might be worth re-iterating this.
* 181-182: Are these references associated with LSTM and Seq2Seq? If so, it would add to the clarity of the paper to perhaps add that.
* **222: Is 4.1.1 included in this, or does subsequent apply to 4.1.2 onwards?** Text has been added to clarify this.
* 229: I think it would also be useful to be clear how many training data were used in each epoch (is it just the number of times you can slide the sliding window for one pixel, or that number x the number of pixels in an interferogram).
* **319-323: The first sentence of this paragraph seems at odds with the remainder.** The first sentence was referring to spatial related stats.Clarifying text has been added.
* **Figure 7: Comparing the seasonal signals and random signals is tricky when they are on different lines/rows, though it is understandable that it is hard to avoid a very wide figure. Perhaps only plotting months 1,2,3,4,6,8 would be sufficient?**
* 355: It may be useful to clarify how this was trained differently to the Seq2Seq models?
* **367-368: I find this section hard to follow. The reported p-values vary by ~10^(33)? Latex error perhaps? These are the correct values.** The linearity is clearly seen in the second two images in figure 9 (giving the high p-values).
* **Figure 9: One of the subplot's title is LSTM8?** Changed to the correct label

**Typos/Edits**

* **118: Apostrophe before Seasonal... incorrect (latex).** Corrected
* **209: Opening quotation mark (latex)** Updated
* **211: Ditto** Corrected
* **273-274: In this pdf, there appears to be an equation missing here, but what appears to be an image of one appears lower down.** This has been changed to a tikz equation rather than an image so that it places correctly in the manuscript.
  + Figure 6: Font may be too small.
* **Figure S4: Legend covers some of the plot.** Updated
  + Figure 8: Font may be too small.

**Reviewer #2 (Formal Review for Authors (shown to authors)):**  
  
The paper aims to use temporal modelling of time series data as tools to forecast  
Deformation in Sentinel-1 InSAR Data. In principle this can be a very interesting approach to improve the forecast of the deformation as it will exploit the temporal dependencies between the data points.  
  
The main weakness of the paper is the technical soundness of the methods, whether SARMIA, or LSTMs. The authors made poor design decisions making the evaluation of the results inconclusive and negatively impacted the potential usefulness of the proposed methods. The paper lacks reasoning behind the choices of the parameters of the models used especially the architectural design of the deep learning models. Another major issue is that the authors target annual seasonality throughout the paper, but the models are expected to learn on 9 months data. The task design of predicting the same length of data as the input features made it unnecessarily more challenging for the models to perform well.  
  
Following are detailed comments:  
1- The paper concludes (in the abstract and conclusion) that "This proof-of-concept study demonstrates the potential of time-series prediction for InSAR data but also highlights the limitations of applying these techniques to non-periodic signals or individual measurement points" This contradicts the purpose and design of temporal modelling in machine learning. LSTMs, for example, try to exploit the periodic and non-periodic information and target mainly modelling temporal dependencies between data points. If the signal is known to be highly periodic (like in Figure 10) then any machine learning technique will be an over-kill.  
2- Line 88-90 statistics about the missing data should be provided. This is important to understand the size and frequency of the gaps in the data. Linear interpolation is only reasonable for small infrequent gaps otherwise this might significantly impact the quality of the model.  
3- Line 116-117: the authors focus on the measuring annual seasonality, but the data is only for couple of years, which means the expected seasonal behaviour is very difficult to capture anyway for lack of seasonal data points. The authors then go on to predict the next 9 months' worth of data which is below their targeted seasonality! A better design would have been to include at least a few occurrences of the targeted seasonality, in this case a few years.  
4- The choice of seasonality given the size of the data is not clear. looking at the data in the frequency domain, using Fourier transform, can give a very good picture of the frequencies in the data which can be an easy tool to find candidates for seasonality.  
5- In line 175, Selecting ACF needs to be justified.  
6- In line 192, Ny is set to 9 months to predict another 9 months?! That is not an ideal design decision. The models are expected to predict the same lengths as the time series they train on. This is going to be difficult for any model to do accurately. The authors did not provide any reasoning for this decision or any analysis to support it.  
7- The models should have been evaluated on the same number of expected future predictions. This is mainly to reflect the error in the loss function used to optimise the model parameters.  
8- Line 206-207. incorrect definition of the models. RNNs actually fail on long sequences, hence the development of LSTM.  
9- LSTM1-4 It is not clear as to what is the size of the input layer. It is important to know this number in order to understand the authors' decision on the parameters of the architecture.  
10- The architecture of the LSTM models is not reasonable. For example, an increase in the number of input features, from one model to another, was not reflected by any change to the number of hidden units or number of layers in the model. I recommend testing different architectures. It is also recommended to optimize the network parameters as much as possible given time/resource availability. It is confusing that the seq2seq models use the same architecture as the LSTM models when they are designed to perform different tasks (from ML perspective).  
11- I generally disagree with the interpretation of the results. For example, line 305 the authors claim the SARIMA and LSTM3 performed marginally better than the constant prediction. That is not how I interpret the figure. The range and IQR are very big for most methods while the medians do not appear to be significantly better than constant. Statistical significance tests should be used where appropriate.  
12- Line 319: the authors confuse the number of features with the size of the dataset. In the presented scenario, they are increasing the number of features not the size of the dataset.  
13- In Figure 7, the IQR/range in the RMSE is quite large for all the models and appears significantly larger than the ones from the constant, which undermines the confidence in the produced results.  
14- Line 359 and 360: it is not clear if the analysis done with re-training of the models from scratch or just testing the already trained ones. On a more general point, however, the use of sequential models is not limited to seasonal/periodic data, but rather to capture the temporal dependencies between data points which might come at different frequencies.  
15- Line 368: use the Pearson's correlation here with the associated p-value. visually there is some correlation for SARIMA and SINU, but it does not look very strong.