I thank the reviewer for his comments and suggestions.   Below I have responded to the comments and suggestions and have indicated where changes or corrections in the manuscript have been made.

My primary concern with this paper is the premise that Method 2 is somehow better for GIC applications because it produces a predicted E(t) that is more consistent (as measured by prediction efficiency and error spectra) with measured data. The problem is that Method 2 has less predictive power than Method 3. Method 2 is subject to time-varying cultural noise that can often be removed using remote-reference techniques. Method 2 is also subject to non-plane wave signals which are often down-weighted through robust statistical processing. The result is that the Method 2 impedances will work well when applied to a time period around that over which they were determined, but generally will perform worse during other time intervals. In short, the differences between Z calculated via the two methods is that method 2 reflects local noise and signal characteristics that change throughout time while method 3 knocks down these effects and represents the plane wave signal. Given the above, I would expect method 2 to perform worse than method 3 when applied to B(t) from some future time interval.

Additionally, the impedances as calculated via the 3 methods need to be shown along with error bars (for methods 2 and 3).

The motivation for using Method 2 was not because I felt it would be necessarily better. As noted in the introduction (and now in the description of Method 2), the motivation is that it is simple and is something that has been used by GIC practitioners in the recent past.   To paraphrase another reviewer, the derivation of transfer functions using Method 3 for the use of conductivity estimation is so complicated that only a MT professional should do it. If this is the case, then we must understand more simple ways of predicting GIC that can be implemented by a non-MT expert when a pre-computed transfer function is not available.

With regard to the issue of performing worse during other time intervals due to cultural noise and non-plane-wave signals, I actually addressed this issue by testing the model developed during one two-day time interval on a different two-day time interval. It seems unlikely that the cultural noise and non-plane-wave part of the signal are of significant concern because the out-of-sample and in-sample prediction efficiencies were similar. In addition, although I have presented the results from four sites in this manuscript, I have studied the data from many others. Based on this, I was confident that the results are generalizable enough that when more sites are considered, the results and conclusions presented here will be similar.

I agree that many of these issues mentioned in this comment are of possible concern, and part of the motivation of this manuscript was to provide a first analysis that attempts to determine if these concerns are significant with respect to the prediction problem. In principle, Method 3 *should* have more predictive power based on many of the arguments that have been made historically in the MT literature and noted by the reviewer.

I do not believe that the intercomparison of the transfer functions are relevant enough for the objective of this manuscript to merit their inclusion. The main objective of this work was to determine out-of-sample prediction performance of different methods for the geoelectric field, not to compare impedance tensors for the purpose of MT/conductivity imaging. Such comparisons have been made extensively in the literature (e.g., Jones et al. 1989). In addition, I have removed a sentence in the summary and conclusion section which mentions their intercomparison.

In conclusion, the comparison between the 1D impedance-based E-field estimates and the 3-D impedance-based estimates is valid and important and should be the highlight of this paper. GIC analyses based upon 1D impedances continue to be used in industry. This makes this work all the more important as it highlights the limitations of the 1D impedance approach. The variation between the two different 3D approaches is not as significant as presented in the paper (for the reasons described above) and should, in my opinion, be given less weight. I recommend that this paper be published following moderate revisions along the lines described above. Specific comments are detailed below.

I believe that the variation between the two different 3D approaches is significant enough for the inclusion in the manuscript because the results presented in this work are not consistent with what was expected, and there is no literature to test that what is expected actually occurs. In addition, it is my preference that when presenting the results of complex models, a comparison should also be performed with a simple model as a base-line. (For example when using a neural network model for prediction, the results should be compared with a simple linear model.) This was another motivation for including Method 2. To address this comment, I have modified the abstract to indicate that the analysis used time intervals where the measurements had few data spikes and no base-line jumps. In addition, I have added a sentence in the revised modified manuscript that notes

"Although the conventional least squares method has been shown to be flawed with respect to transfer function estimation for the purpose of ground conductivity estimation (Egbert and Booker 1986), we have shown here that it can produce equal or improved out-of-sample predictions of the electric field on data segments without many defects."

P1, L10. 'bias-corrected' is unclear and not really described ever. The remote reference and the statistically robust parts are clear and well defined in the references that describe these techniques, but not the bias correction part.

I have removed the "bias-corrected" part of the sentence as it was redundant given that I state that the method uses remote reference measurements.

P2, L23-27. The distinction between the two items described here (either an effective approximation of 2D/3D conductivity structure and an effective approximation that produces reasonable GIC estimates) is unclear.

I have modified the locations of the references in the sentence to clarify this. The sentence now reads

"... conductivity structures that may exist (Fernberg 2012) or an effective approximation that may not reflect conductivity structures but produces reasonable GIC estimates (Boteler 2015)."

Fernberg 2012 developed 1-D layered models of conductivity for the purpose of approximating structures that may exist; Boteler 2015 uses several layered models for GIC estimates but notes that although they may not reflect the real conductivity structures, they can be used for reasonable GIC estimates until better models are available. The distinction is in the motivation - Fernberg 2012 was concerned with obtaining approximate 1-D models that are consistent with physiographic geologic structures whereas Boteler was less concerned with physiographic geologic consistency and more with a model that produced the best GIC predictions. (In an early draft of this manuscript, I did not make this distinction and a GIC practitioner was insistent on making the distinction.)

P2, L44-47. I'd suggest changing the order here to have error bars (a statistical metric) before visual characteristics (a subjective metric). Also describe what you mean by 'consistency when different "data segments are used...' I presume you are referring to the mathematical stationarity of the response (doesn't change when computed using a different time window).

I have changed the order so that "error bars" is first.

I have modified the sentence from "and consistency when" to "... and consistency of the computed transfer function when …".

P3, L58-68. This reads like you are concerned with how Z is estimated. The abstract, however, seems more concerned with the complexity/dimensionality of Z. As a note, there have been numerous papers throughout last couple decades comparing various ways to calculate Z, albeit for the purpose of MT/conductivity imaging.

The manuscript is concerned with both. Method 1 was used because it has been used historically in the GIC literature. Method 3 was selected because of the availability of the precomputed transfer functions. Method 2 was selected because it was simple and from experimentation I found that it quite often produced superior out-of-sample prediction efficiencies than Methods 1 and 3 and because it has been used in the GIC literature. In addition, as noted in the introduction, Method 2 has been used by GIC researchers. Initially I had considered also using a 1-D or 2-D version of Method 2, but decided that there was not much of a point as the computational and implementation complexity is similar to that of the 3-D version. To address this concern, I have added an additional sentence in the last paragraph in the introduction:

"That is, Method 2 has been used in the past for purposes of GIC estimation and Method 3 has been used in the past for estimating ground conductivity structures; in this work I compare both methods with respect to electric field estimation (which is used for GIC estimation).".

Regarding papers that compare methods for calculating Z, the original manuscript cited Jones et al. 1989 several times, which compares six methods of calculating Z (but does not compare the ability of Z to predict out-of-sample E as was done in this manuscript).

P4, L77-78 and eqn 2. Need to explain in a few more sentences. Z_N(omega) is not defined. It should be noted that this is a recursive analytic solution.

I have modified the sentence to be “A surface impedance, Zn, is computed ...” instead of “A surface impedance is computed ...”.

I have modified the sentence from “from the use of Wait's recursion formula” to “from the use of Wait's analytic recursion formula”

This part of the manuscript is intended to only be a brief summary of the methodology and I gave two references for details on how the calculation is actually performed. Providing sufficient detail for the reader to reproduce the calculation would require a significant expansion of the section and I felt that it is better for the reader to refer to the references so as not to distract from the main objective the manuscript.

P4, L79-80. How are you carrying out the inverse FFT? The details of how you are conditioning your spectra prior to the iFFT are not described. Also, are they consistent between the three methods?

The inverse FFT was carried out by the convolution of the estimated Z (interpolated onto a uniform frequency grid) with the raw spectrum for B. The spectrum of B was not conditioned. I have added a sentence to indicate this in the description of Method 2.

“The spectra of Bx(w) and By(w) used in the inverse fourier transform was not pre-conditioned. The results were insensitive to the method used for interpolation of Z (i.e., cubic interpolation or interpolation in log space).”

P5, L86-90. This is a confusing description

I have added a reference to Equations 4.17 of Simpson and Bahr, 2005 and modified the sentence to be

“In this work, the evaluation frequencies were selected to be logarithmically spaced (as described below) and the auto-- and cross-spectral values required for computing the elements of Z (Equation 4.17 of Simpson and Bahr 2005) at each evaluation frequency are determined using a Parzen averaging window on the raw spectra.”

P5, L86. Suggest replacing 'largest' with 'highest.'

Done.

P5, L91. 'linear interpolation.' Specify whether you are doing linear interpolation of the complex Z values or separately on the real and imaginary components. Should be doing the former.

The linear interpolation was done separately on the real and imaginary parts and I have added this statement to the sentence referenced. I agree that for purposes of conductivity imaging, interpolation should be done on the magnitude and phase as they are the quantities of interest for physical interpretation of the transfer function.

P5, L95-100. Why all the machinations here?

The point of the paragraph was to document the experimental steps and observations that were made in doing the analysis. In discussing how to do some of the computations with MT researchers, and reading the literature, I have found that many processing steps are stated as being quite important, but quantitative justification is difficult to find. As an example, I was told by a MT researcher that the use of a Parzen window was "very important" and that a rectangular window should not be used, but could not find in the literature any evidence to support this.

P5, L98. 'slight improvements.' In what? What is your metric here for improvement?

I have modified the sentence to include "in the prediction performance". The sentence now reads "With this, the use of a Parzen averaging window provided slight improvements (~2 %) in the prediction performance over that for a rectangular or Bartlett averaging window."

P6, L113-114. '...and also had four day time intervals...' Say what you are talking about here - MT data at the particular station? mag observatory data?

I have modified the sentence to be "is available and also had four-day time intervals of **E**(t) and **B**(t) measurements with few spikes ..." instead of "is available and also had four-day time intervals with few spikes ..."

P6, L118-120. Provide a sentence on how you did this.

I have modified the sentence to read "Data spikes were manually identified and replaced with linearly interpolated values" instead of "Data spikes were manually removed".

P6, L121-123. What are you saying here? Unclear. Explain.

I have replaced the sentence with

"The motivation for the zeroing of frequencies outside of this range is to allow for a comparison the prediction performance of all three methods with impedance tensors that span the same period range."

P7, L151-152. As the difference between the two methods is dependent upon the zeroing out of the periods this really needs to be explained.

At the end of the paragraph I have replaced the existing sentence

“This is explained by the fact that frequencies outside of the range of 9.1 − 18, 725 s are predictable for this site.”

with

“This small improvement for Method 2 is explained by the fact that periods outside of the range of 9.1-18,725 s are slightly predictable for this site. Because Method 3 does not produce predictions outside of this period range, its overall prediction performance decreases because of the increased variance in the measurements, which increases the denominator in the ARV.”

P7, L 156. '...the frequencies outside ...are predictable...' Are you actually calculating them? Typically the Method 3 impedances are calculated to the shortest stable period possible given the sampling rate and to the longest stable period possible given the time series length and signal levels. How are you recovering shorter and/or longer periods?

I simply changed the allowed frequency range of the impedance tensor computed for Method 2 to be the maximum range allowed by the sampling frequency and the length of the data set.

The cut-off periods for the Method 3 impedances are determined in part by a cut-off in coherence. As the coherence is a measure of predictability, if one considers a lower cut-off in coherence, corresponding to widening the period range of the computed transfer function, one is expected to have higher prediction efficiencies.

Said another way, for MT/conductivity imaging, the cut-offs are determined by a criteria associated with the quality of the estimates of the transfer function. Here I am asking if we relax that criteria for the purpose of prediction, can we get better predictions.

P8, L164. 'The smoothed error spectra...' Note explicitly that you are applying this method to the error time series shown in figures 2-5. Otherwise it is not clear what you mean by error spectra.

I have modified the sentence to state “The smoothed error spectra in Figure 1 for the time series shown in Figures 2-5 were ….” instead of “The smoothed error spectra in Figure 1 were ...”

P8, L180-181. This could be emphasized by stating it simply. The errors for the 1D approach are larger than the measured data itself.

I have modified the sentence to say … “has errors that are less than the measured amplitudes” instead of “has error amplitudes that are less than the measured amplitudes”. I chose not to replace “measured amplitudes” with “data” to avoid ambiguity.

P9, L200. '(1-D assumption)' this is only part of the 1D assumption. Also that the diagonals elements are zero.

This sentence has been modified from “(1-D assumption)” to “(part of the 1-D assumption)”. (For the sites considered, the diagonal elements much smaller than the off-diagonal elements. This is the reason that I do not mention ignoring of the diagonal elements as being a primary reason for Method 1 producing inferior estimates.)

P9, L210. '...when making unbiased estimates...' Also stationary (time invariant). This is critical in order for the impedances to have any predictive power beyond the immediate time interval used to predict Z.

I addressed the stationarity issue by computing the transfer function for one two-day interval and then used that computed transfer function to compute prediction efficiencies on a separate two-day interval. If the signal was highly non-stationary on this time scale, one would get very different prediction efficiencies between the in-sample and out-of-sample tests.

P10, L212. '...in practice, remote reference data...' If you don't have remote reference then you automatically would be doing a variant of method 2. If you apply method 3, however, it is superior to 2 in that it is good for all time.

Indeed, it would be a variant of Method 2, but there are still many additional processing steps and assumptions that make Method 2 and Method 3 distinct. For example, the quantity minimized for Method 3 is a weighted least squares of the error and there are various methods for determining the weights.

P10, L217. '...from Method 2 differ from those of Method 3...' Show the reader the impedances with error bars from both methods. How do they compare?

I have removed that sentence based on a comment by Reviewer 1. As noted in that reply, the intended meaning was that if we don't know the ground truth, we can't say which model is better. In retrospect this seems somewhat obvious. I agree that there is a vast literature on why the model associated with Method 3 is *expected* to be closer to the ground truth.

P10, L226-228. The underlying data used to compute method 3 has no greater spectral content. This is a limitation of the data used to compute Z, not the method used to compute Z.

The underlying data has a cadence of 1 second and the spectral content in the range of 2-9 seconds is non-zero for both **E**(t) and **B**(t), so Z can be computed in the range of 2-9 seconds. As noted earlier in this reply, in MT studies, Z is not computed in this range because of the cut-off criteria for the coherence.

Figure 1 - at longer periods the errors increase because the Z estimates are worse and you start seeing the effects of non-plane-wave effects.

I agree that this is may be true, but without additional analysis that includes consideration of the ionospheric structure during the time interval considered, I do not want to commit to this conclusion. It is important to note that if this were a strong non-plane-wave effect, one would not have obtained such similar in-sample/out-of-sample prediction efficiencies (unless the non-plane-wave effects happened to be identical between the in-sample and out-of-sample intervals considered, which is unlikely because of the similarity of the prediction efficiencies between the two intervals for all four cases considered and the fact that non-plane-wave structures vary on time scales much shorter than two days).

Figure 1 - put all subplots on the same scale to allow for direct comparison.

Because the objective of each plot is a comparison of the lines within each panel, I put each panel on a different scale to highlight the intra-panel differences. In the manuscript, I do not describe anything that would merit having all subplots have the same vertical scale.