Response from the authors to the comments by anonymous referee

We would like to thank the referee for providing constructive review and commentary.

I very much appreciate to investigate the numerous options of ML for geophysical application and novel ideas related to this issue are of particular interest for HESS. In this manuscript ML was intended to be used to improve a design optimization task for electromagnetical field mapping. The approach is interesting and especially the interpretation of the feature importance has an added value as this allows some enhanced interpretation.

We are happy that you appreciate the investigation of machine learning for geophysical application and find our approach interesting. We agree that the interpretation of feature importance is the most interesting nugget!

The manuscript holds a lot of interesting results however I suggest to rethink the focus of the manuscript. In the recent form of presenting the methods and results I cannot agree that "The result is an approach that can allow an EMI user with limited expertise to choose a better set of instrument configurations given their main survey goal and knowledge of the site conditions. (line 493/494)".

One of my concerns is that the authors formulate as their main objective to present an approach to select sets of EMI configurations that are optimal given the specific survey goals and any independent knowledge of the subsurface electrical properties - with the aim to support users with limited expertise, see line 67-74. To fulfill this aim it would be more helpful to write a practical guideline than a scientific paper. In the recent form I have doubts that the manuscript can support users with limited expertise as the figures and way of recommendation needs to be simplified.

Thank you for finding our results interesting. We agree that we should step back from the goal of making a simpler approach and redirect the focus towards the scientific value of the study. We have refocused the paper significantly based on the reviewer's recommendations and greatly appreciate their perspective. This has led to a fundamental change in the objective of the paper that we find much more compelling – again, we thank the reviewer for their insight.

Moreover the authors choose a rather arbitrary selection covering a very broad range of subsurface properties for the forward models. The chosen ECa range is rather high and from the practical point of view many field sites vary by a delta ECa not more than 20 mS/m which would cover only two classes (e.g., van Hebel 2018, McLachlan 2017, Robinet 2018, Reyes 2018).

The full ranges of the subsurface properties are supposed to cover the range of many areas. This is to simulate a scenario where the same user must survey multiple areas that not necessarily similar and we therefore consider a wide range of geology, which can have a large variation in EC (Palacky, 2011). This could apply to an investigator that is tasked with surveying multiple fields but wanting to keep the design the same for intercomparison purposes, or who is conducting a survey over a large or rather heterogeneous area.

However, our later analysis shows how a user can choose to only consider a narrow range of values if the site conditions are better defined. When we constrain the subsurface ranges in section 4.4 and 4.5 it is to illustrate that there can be a benefit to changing the instrument setup based on the specific field. Figure 1 shows the ECa measured with a horizontal coil at 2 meters separation. On this field the range of ECa values varies from 1.6 mS/m to 99.3 mS/m. While this kind of variation might not be the norm, we left in the possibility that it can occur. In addition, the approach could be constrained to consider high resolution

within a narrower range of EC values to give a user insight into how finely EC could be constrained with EM instruments.

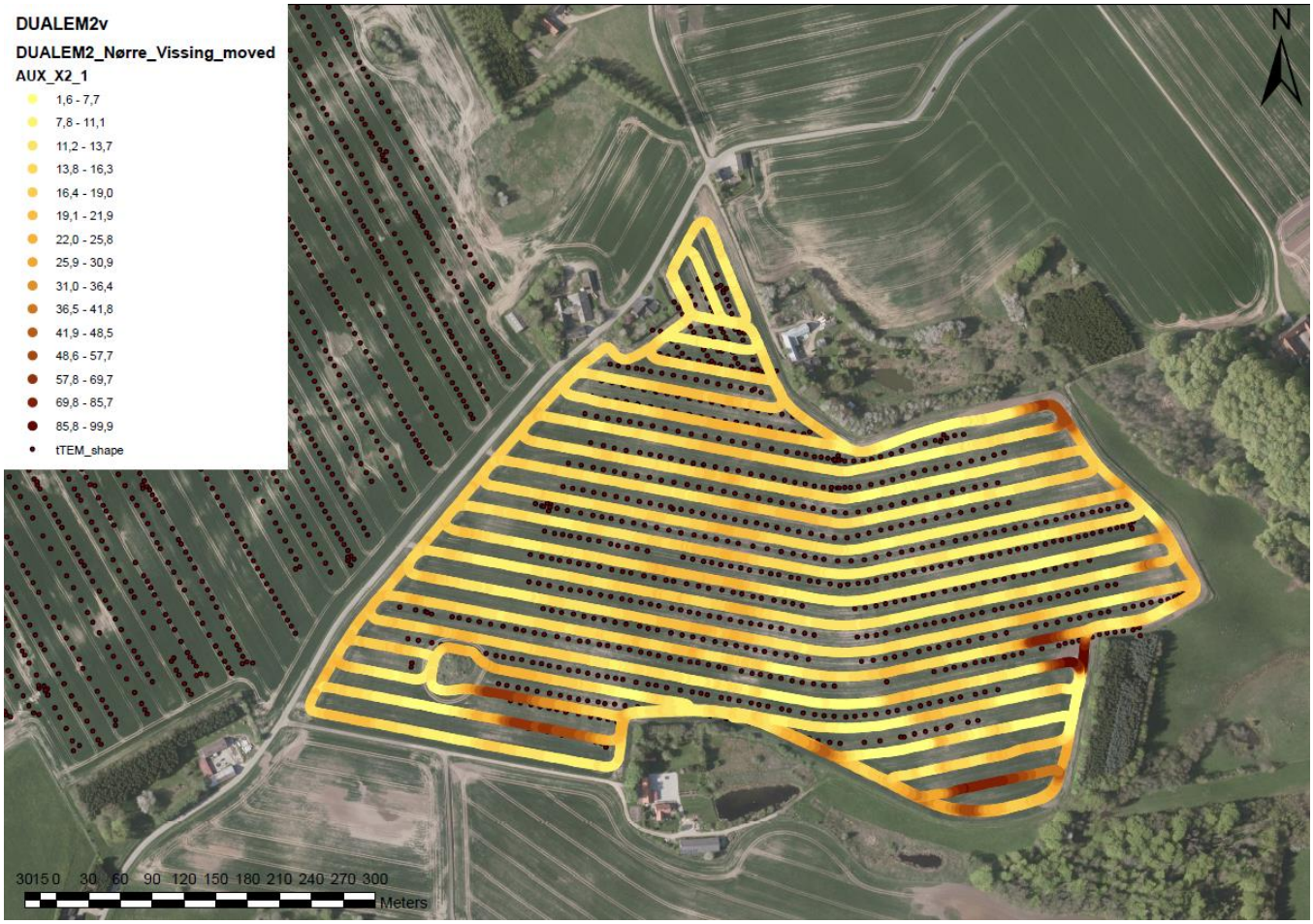

Figure 1 Raw ECa measurements from the horizontal coil with 2 meter separation in a dualem21 instrument. The field is located at coordinates 56°07'40.3"N 9°51'45.0"E in the central Jutland, Denmark.

Given the option of EMagPy it seems to me more convenient, even for an unexperienced user, to run a forward model with several instrument configurations (HCP, VCP, PRP and coil distances) for the specific application with some prior knowledge of texture, salinity etc..

The purpose of this approach is to reduce the bias that comes from the suggested approach. How does a user decide on which small set of configurations to consider? How do they quantitatively compare the likely success of these proposed configurations? Our idea is to provide a simple, objective approach that can explore many possible configurations – including some that may not be in popular use. Furthermore, even considering only a few configurations, the user would have to consider multiple combinations of these configurations, which quickly becomes impractical. If we then include a sensitivity based on existing knowledge the number of simulations can become huge and interpretation requires more effort than most investigators will commit. (Perhaps this is one reason that so few pre-survey analyses are conducted to optimize data collection.) We choose an illustrative example of using layer EC and thickness as prior knowledge. But any information could be used to constrain the range of cases that is considered by the machine learning.

Moreover I see a big challenge for unexperienced users to understand the dynamic aspects of the depth sensitivity of EMI depending on the subsurface EC distribution. In this manuscript this aspects was excluded as stated in line 58-59/line 120. I can understand to keep the situation in a first attempt simple in terms of using McNeill model, however I would strongly avoid to make decision on measurement configurations without keeping this aspect in mind.

This is a good point. Fortunately, because EMagPy includes forward models that consider (to some degree) the impacts of conductivity structure on the EMI response, this would be a trivial extension. If the conditions warranted the added effort (i.e., the LIN assumptions are clearly violated), then the user could implement an even more complete forward model within the ML structure shown here; the only cost would be the forward model run time. Our choice to use McNeil was based on two things. First, McNeil is still the most widely used model for interpreting EMI data – we contend that the data collection should be chosen with consideration of how the collected data will be analyzed. Second, we wanted to make the connection between ML recommendations and underlying concepts. For a broader audience, we felt that these discussions would be clearer if based on the relatively simple cases for which McNeil applies. (If the reviewer is interested, we discuss why a more complex forward model actually provides even greater advantages for our proposed approach compared to traditional inverse model approaches to data worth analysis in response to the other reviewer's general comments.)

My suggestion would be either

- to focus on a very practical guide for users based on forward modelling that not only includes the instruments configuration but also EC of the subsurface and including a real world example to transfer knowledge into practice

- or to focus on the scientific value of the study and rather present and discuss your approach (and its advantages) compared to existing approaches/forward modelling having more room for a structured discussion (e.g. Table 2, Figure 4 and Figure 7) and advancing the way of presenting the results (Fig 7, 8). Especially for the results in chapter 4.4 I do not see the added value clearly.

We appreciate the reviewer's advice. We have significantly refocused the paper on the scientific value – how ML can provide an objective approach to assessing the likely information content of a wide range of possible measurement sets. However, we have maintained some extension of the work into practical implications because we feel that EMI is, ultimately, a highly applied method more so than a research-grade instrument.

We see the value of section 4.4 analysis as providing a quantifiable way of assesing how well an EMI survey will fare depending on the goals and and field conditions of the survey. Rather than depending on a rule of thumb (see below). The change in NRMSE creates a measure of how idientifiable a parameter is. Instead of suggesting that thin layers are hard to detect we can quantify how much harder they are to detect and at what thickness it becomes impractical to use EMI.

We now have explicitly defined the general rule of thumb in the introduction:

"The depth of investigation (DOI) of EMI instruments is both in the scientific literature (Saey et al., 2009a; Saey et al., 2009b; Saey et al., 2012; De Smedt et al., 2014; Doolittle & Brevik, 2014; Adamchuk et al., 2015) and by the manufacturers (Dualem Inc., Canada n.d.) often estimated to be at the depth the has 70% of the cumulative response. There is a relationship between depth sensitivity of the instrument response and coil

spacing and position. Therefore 70% cumulative response rule is in practice frequently converted to a rule of thumb that states larger coil spacings and HCP should be used for deeper investigations while short spacing and VCP/PRP should be used for shallow investigation (Acworth, 1999; Beamish, 2011; Cockx et al., 2009; K Heil & Schmidhalter, 2015; Kurt Heil & Schmidhalter, 2019). While this rule of thumb is not wrong, the terms shallow and deep are subjective and will have different meaning depending on whether it is a hydrogeologist, archeologist, agronomist or a geophysicist who applies the terms. It also fails to make any distinction to the differences between using the VCP or PRP coil orientations."

We edited the aim to:

"One of the challenges of both scientific and environmental investigations is to determine the optimal data to acquire. Data, which is often used to provide structural information to a model or constrain model parameterization. Measurement optimization is an attempt to balance data quality and the work expended in the field and laboratory. The ultimate goal of was to develop a robust approach to measurement optimization, with the hope that a similar approach could be extended into other measurement network design problems."

Specific comments:

- in the title the root zone is explicitly mentioned however it doesn't appear later on to be an issue

We will change the title to

"Using Machine Learning to Predict Optimal Electromagnetic Induction Instrument Configurations for Characterizing the Shallow Subsurface"

- in the introduction you use the formulation "near surface hydrogeologic structure", later you switch to layered soils – maybe you can unify wording

We unified the wording to only use layered soils and changed the sentence to:

"Water movement through the vadose zone is often controlled by the near surface layering of soil."

- the introduction contains many information that are rather a methodological description of your work, e.g., line 57-58, 85-107, please address these issue in the methods chapter

We agree to move the description from l57-58 to section 3.1 and the initial explanations of machine learning (l85-107) to section 3.2. The following remains in the introduction to introduce the concept:

"Machine Learning (ML) describes a wide range of regression algorithms used for pattern recognition. ML has grown in popularity and is now used regularly within and beyond science. The simplest ML tools are based on Decision Trees (DT), which are supervised ML techniques that perform classification or regression by sequential categorization based on observations. DTs are computationally inexpensive, but they can have limited predictive skill (Hastie et al., 2001). To improve their performance, DTs are often augmented by ensemble learning methods such as bagging (Breiman, 1996) and boosting (Friedman, 2001)."

And the following is moved/added to method section 3.2

"We found that gradient boosting (Elith et al., 2008; Friedman, 2001) offered improved performance without adding unreasonable additional computational effort and it was used for all analyses. For our application, each modelled EC_a value in the ensemble of the different EMI configuration represents a

feature in ML parlance. We then tested the ability of DT with GB to infer the correct value of each subsurface property given the EC_a that would be measured with all the EMI configurations."

"We used the feature importance capabilities of DT with GB to identify which observed EC_a values were most informative for the inference and eliminated all insensitive configurations. This allows us to find the optimal instrument configurations for each subsurface parameter without having to do inverse modelling. To examine the impact of independent knowledge of any of the subsurface properties, we then repeated this analysis for a subset of the soil models that met a given restriction, such as only those that had a thin upper layer or a high EC middle layer."

- In order to simplify your discussion and figures the height above ground could be released in a first step, since the assumption the all option are in any case available is misleading, e.g., I don't think its possible to carry an instrument with a coil distance of 4m at a height of 10 cm above ground along an agricultural or grassland transect. I completely understand that it is tempting to use all the information since ML is designed for big data, however for better understanding you could make use of Fig.2 in combination with some practical issues to reduce input heights.

The Department of Geoscience at Aarhus University has a Dualem421S system that can be towed behind an all-terrain vehicle (https://www.aarhusgeoinstruments.dk/dualem). While most fields are not completely leveled, the towed instrument still secure uniform instrument height that is close to uniform. In addition, there seems to be persistent interest in making measurements at multiple heights (e.g. ground placement and hip height) to improve information content.

- Do you have an idea why is the residuals in Fig 3 and 6 not evenly distributed? low EC values are overestimated and high EC values are underestimated - this aspect of heteroskedasticity needs to be discussed

We would argue that the skew is relatively small and limited to the extreme high and low values. Most of the residuals are symmetric. To explain the extreme values, we expect that this is due to the limits on the input values of 0 mS/m to 100 mS/m. Therefore, as the true cases approach this limit there are no EC values below the minimum (above the maximum) that can provide symmetric residuals.

- Fig. 4 I agree that a problematic condition for EMI is the thickness of a layer which is shown nicely for the thickness of A – the thickness of B should be even more challenging however this is not represented in the "outliers"

Fig 4. Shows the distribution of values within the outliers (1 std. off) from fig. 3. ECA is the parameter that is being inferred and the distribution of thickness A values in the outlier set show that small values of thickness A are dominant. While the value distribution of thickness of B is uniform and therefore no specific thickness of layer B makes a worse inference of ECA.

Fig 4. Could be reproduced for each of the five subsurface parameters, but we have chosen to only do it for ECA as is also the case with fig 3. and 6. This is partly because the same information is presented later (Fig. 7), but for all parameters instead.

Figure 2 of this document is Fig 7 from the manuscript. Here the center column represents the attempts to infer ECB. Changing the range of thicknesses of layer B is shown with gray markers. It is worth noting here that the modification of thickness B provides the most dramatic differences in NRMSE between the three

restriction patterns. With a high NRMSE for the thinning of the layer (triangle) and a low NRMSE for the thickening (square) of the layer. The high NRMSE for the thinning is larger when inferring ECB than for inferring ECA showing that a thin ECB is even more challenging to detect, as the reviewer surmised based on their experience.

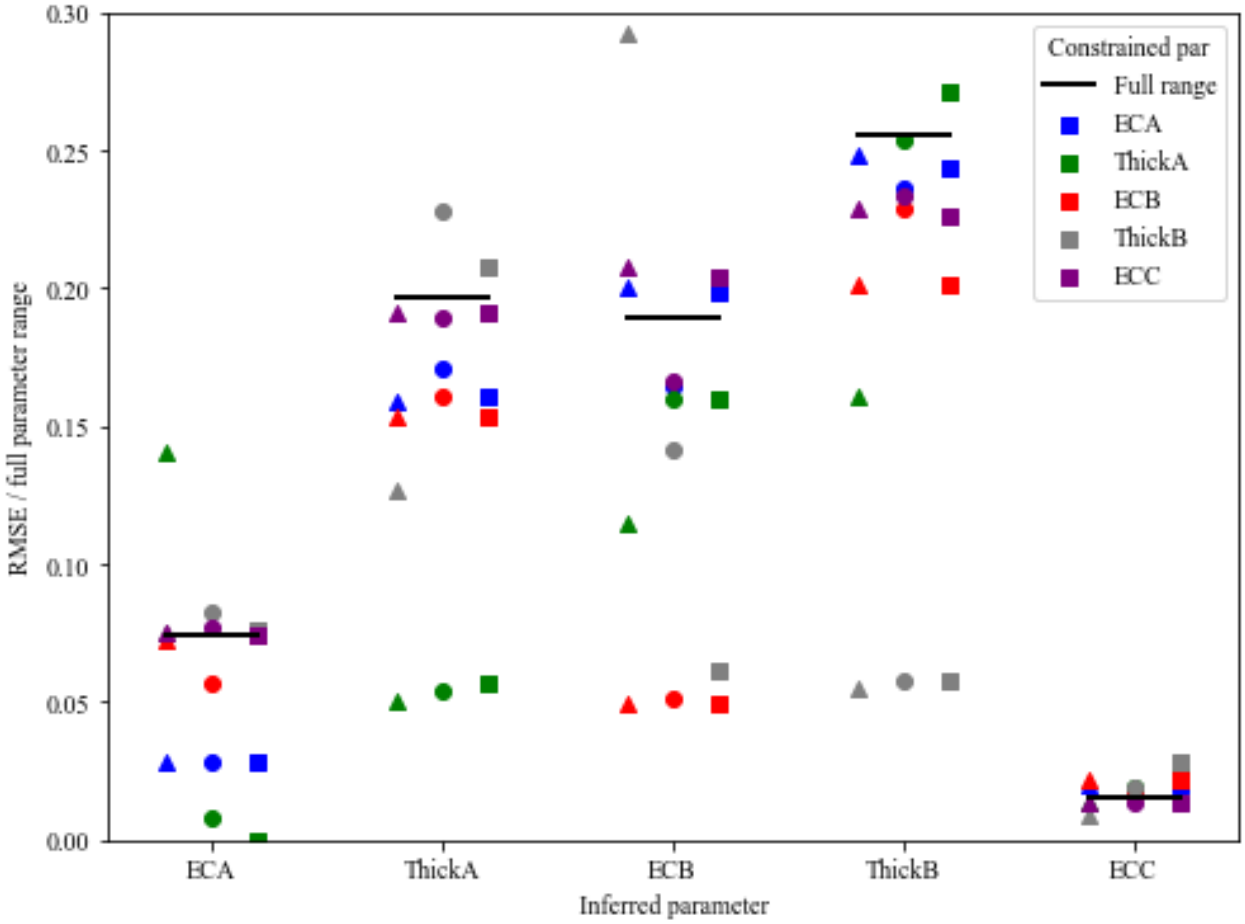

Figure 2 Figure 7 from the manuscript with a caption that reads: "The changes in inference of the five subsurface parameters (x-axis) are based on a comparison between the RMSE from restricted case divided by the range of the parameter (Y-axis). The lines show how well the parameters are predicted when all parameters are full range. The color shows which parameter that is being represented and the location and symbol represents the three restriction patterns skewed low (left nudged triangle), centered (centered dot), skewed high (right nudged square)."

- the usage of an NRMSE is not clear to me if you intend to guide the user directly (l468)

We have edited the manuscript to make sure that the term, NRMSE, is clearly defined and to state simply that the use of NRMSE is to inform the user if the change in instrument setup will provide higher quality data (lower NRMSE) or lower quality data (higher NRMSE)

We added the following paragraph to the section 4.2 where NRMSE is first mentioned.

"The NRMSE of the parameter is a measure of how well the ML is able to infer the individual parameters and thus how estimable the parameters are. Because the ML is trained on EMI output the NRMSE also suggests how well the EMI instrument can detect the soil properties"

And also added the following at the specific line (468) you refer to:

"A low NRMSE will suggest a more reliable characterization of the subsurface property by the instrument and vice versa."

Referee references

McLachlan, P.J., Chambers, J.E., Uhlemann, S.S. & Binley, A. 2017. Geophysical characterisation of the groundwater–surface water interface. Advances in Water Resources, 109, 302-319.

Reyes, J., Wendroth, O., Matocha, C., Zhu, J., Ren, W. & Karathanasis, A.D. 2018. Reliably Mapping Clay Content Coregionalized with Electrical Conductivity. Soil Science Society of America Journal, 82, 578-592.

Robinet, J., von Hebel, C., Govers, G., van der Kruk, J., Minella, J.P.G., Schlesner, A., Ameijeiras-Mariño, Y. & Vanderborght, J. 2018. Spatial variability of soil water content and soil electrical conductivity across scales derived from Electromagnetic Induction and Time Domain Reflectometry. Geoderma, 314, 160-174.

von Hebel, C., Matveeva, M., Verweij, E., Rademske, P., Kaufmann, M.S., Brogi, C., Vereecken, H., Rascher, U. & van der Kruk, J. 2018. Understanding Soil and Plant Interaction by Combining Ground-Based Quantitative Electromagnetic Induction and Airborne Hyperspectral Data. Geophysical Research Letters, 45, 7571-7579.

Author references

Palacky, G. J. (2011). 3. Resistivity Characteristics of Geologic Targets. *Electromagnetic Methods in Applied Geophysics*, 52–129. https://doi.org/10.1190/1.9781560802631.ch3