Earth Syst. Sci. Data Discuss., referee comment RC2
https://doi.org/10.5194/essd-2022-297-RC2, 2022

Comment on essd-2022-297

Anonymous Referee #2

Referee comment on "GlobalWheatYield4km: a global wheat yield dataset at 4-km resolution during 1982–2020 based on deep learning approaches" by Yuchuan Luo et al., Earth Syst. Sci. Data Discuss., https://doi.org/10.5194/essd-2022-297-RC2, 2022

General comments

This study leverages phenology-based mapping of wheat spatial distribution in conjunction with both ML (RF) and DL (LSTM) modelling and global gridded weather data to produce a global wheat yield dataset at 4km spatial resolution from 1982 to 2020. The manuscript is generally good, though major revisions are required before recommendation for publication.

Specific comments

- Title – Deep learning is a type of machine learning. Correct the title to read '…based on machine learning approaches' because this is accurate.
- 43 – ML approaches are a form of statistical model, so cannot be an alternative to statistical models. Could phrase as 'ML provides an innovative approach to statistical modelling and can address…'
- 45 – Kang reference MISSING, add reference to bibliography, then rephrase so that 'statistical models' and 'ML models' aren't seen as separate things, ML is a form of statistical modelling.
- 50 – crucial point, provide slightly more detail about the studies referenced, i.e. what crops/locations was the LSTM performance better than ML?
- 53 – if you're saying 'it is well recognised that…' then at least two references supporting this statement are required.
- 71 – consider reorganising these sections because having Data and Methods as sub-sections is unusual. For example, remove the joint Data and Methods section, have separate sections for each and include the study area in the Data section.

- 98 – Table S1 should include more detail about the sources of information, this is key to the paper and the authors should consider including it in the main manuscript. Additional detail should include how the yield information was collected (farmer reports of area with observed production, farmer yield reports etc) and what organisation collected it (government ministry, NGO, UN etc) because this forms the crux of the dataset
- 109 – Expand this sentence to a brief paragraph describing the overall flow of methodology of the paper, to signpost the reader so they know what sections to expect throughout the methods section. Currently the methods section jumps around a bit and is unnecessarily confusing to the reader.
- 134 – what were the optimum hyperparameter values after tuning?
- 156 – can you clarify if this is out-of-bag RMSE for the RF? If so, please state clearly and briefly explain in the text
- 165-166 – this is a clear example of why the use of 'statistical data' to mean 'observed data' is confusing throughout this paper. Please change all references to country-reported, observed yield data to 'observed data' and remove references to 'statistical data' because it is confusing to the reader when your new dataset has been generated using statistical models. Especially confusing also on lines 122-124
- 173 – give RMSE of areas as percentages of country area rather than absolute values as these aren't relevant when comparing between countries
- 177 – clartify uncertainties o remote sewnsing prpoducts
- 185 – were there any regions in which RF outperformed LSTM? Means of 0.72 and 0.64 are not that far apart and only regions where LSTM outperformed RF are reported, please make it explicit if RF did not outperform LSTM in any regions.
- 188 – wherever you report R^2 values, please also report the associated RMSE or OOB RMSE
- 242 – move uncertainties section into the results section
- 251 – go into more detail about observed yield data availability limitations – how did you overcome them and what were they precisely? Consider building into a new version of Table S1

Technical corrections

- 27 – 'climate variability, extreme weather events and global crises…'
- 29 – 'pandemic is estimated to have added…'
- 38 – 'In addition' doesn't work here, remove entirely or substitute with 'On the other hand' or similar
- 53 – reword in positive manner – 'although there are a few studies…there is still significant development to be done.' or similar.
- 58 – incorrect usage of 'hamper', replace with 'limit' or similar
- 73 – remove pluralisations of area and production
- 118 – incorrect grammar 'when applied it in', please correct
- 133 – samples plural

- 186 – this is the first use of nRMSE and it is not defined (I know you defined RMSE but what is nRMSE?)
- 238 – ‘regardless’ instead of ‘despite’, depluralise years and regions