Comment on essd-2021-442
Anonymous Referee #2

ESSD paper review:

This study describes a new 1km dataset of monthly-mean, monthly-maximum and monthly-minimum surface temperature's over China, developed using machine learning methods. The method used for the final data set was chosen as the best performing method, after a comparison of three modern techniques. A dataset of 613 weather stations over China was used to train and test the machine learning methods. This study is very clearly written and the Figures are of high quality. I agree with all of reviewer 1’s comments, so will not repeat these points and assume they have been addressed within the manuscript, but I will add a few further comments below.

General comments:

- There is always a tradeoff between spatial and temporal resolution when designing new data products. Can you explain in a few sentences in the manuscript why you chose to create a product with such high spatial resolution but such low temporal resolution? You make comparisons at the end to the ERA5 dataset which does have much lower spatial resolution (~30 x less) but it has hourly temporal resolution (720 x more) which is very useful for a number of applications. Comments suggesting the applications where you think this dataset may be preferable to the others mentioned would also be useful.
- You mention ERA5 is only available from 1979, but it is now available back to 1950, so could be used to incorporate dynamical variables (as suggested by reviewer 1). I’m not suggesting you do this, but in the limitations this could be a point for future development. And the text should be updated to reflect the availability of ERA5.
- Is any quality control performed on the meteorological station data you use as inputs? A few stations with low quality data could skew the results in data sparse regions.
- Do you know if the final model output is sensitive to the choice of stations used in the
test/training dataset? I imagine that this could heavily influence the results in the data sparse regions.

- Although you’ve included elevation, latitude and longitude there are multiple climatic regions in China, and a large amount of external drivers to variations in temperatures. The strength of these may modulate surface temperature behavior (e.g. the strength/location of the monsoon circulations, El Nino southern Oscillation, and other global teleconnections). Distance from the ocean could also play a role. Have you considered these in your explanations for months/stations with particularly large residuals, or stations with strange behaviors? It could be that if a month had anomalous large scale weather conditions, which your machine learning methods are not trained to capture there are large residuals? These could make interesting case studies and could motivate future work incorporating some dynamical predictors.

- Figure 2: This is a nice depiction of the relationships. If you could briefly unpack the meteorological understanding behind this in the text it would be beneficial to readers. Have you checked that the relationships hold if different climatic regions of China are subset out?

- Line 190-195. So you have 840 different models. Can you comment on how different are all the 70 models for each month? (e.g. do all the January models look very similar?) This could be useful to understand if there are dynamical meteorological explanations for any outliers.

- Line 243: Do you have a sense of why the errors are larger in the colder months? Are the impacts of local meteorological conditions larger in the cold season, which would make it more difficult for the methods to work? There may be meteorological literature on this.

- Figure 6: Can you comment on any features you’re resolving here that are not seen in the lower resolution gridded products you will compare to? There are some very high resolution features on the map, but clarification that they are physical would be useful.

- Does your trend analysis agree with the existing literature on global warming over China? If so include references to this.

- Figure 11: The Taylor diagrams show clear improvement from your new dataset. Also including some timeseries from locations not sampled from the observation network compared between the three datasets would be useful to understand how the four products sample the seasonal cycles of the variables.

Small corrections:

- The height of the air temperatures (surface, 1.5m, 2m) should be added to the manuscript when this is mentioned.

- The acronyms for datasets/methods should be defined in the abstract to make it easier to read.

- Line 38: after commenting on the limitations of the observing stations you could comment here on the limitations of reanalysis based products.

- Throughout the text when you say ‘high resolution’ this should be changed to ‘high spatial resolution’ e.g. line 55.

- Line 56: ‘traditional interpolation techniques’ might be clearer?

- Line 58-60: You comment on a few studies which talk about the superior performance of machine learning techniques but you do not say what the benchmark is that they’ve
succeeded against. This should be included.

- Line 61: ‘estimation of short-term air temperature’ – I’m not sure what you mean by this?
- Line 86: The link here gives me an Error 404.
- Around the discussion for Figure 1 it would be interesting to know the spatial distance between observation sites. This might be a small indication of confidence in the final machine learning model output.
- Section 2.3: When commenting on the spatial resolution of the gridded products used for comparison it would be useful to also have this in km.
- Line 149: ‘machining learning’ should be ‘machine learning’
- Section 3.2.2 Are the choices of parameters for the SVM method standard in the literature? Can you please comment on your choices?
- Line 342: ‘shows a cyclic pattern’ might be clearer.