Supplement of Earth Syst. Sci. Data, 14, 5037–5060, 2022
https://doi.org/10.5194/essd-14-5037-2022-supplement
© Author(s) 2022. CC BY 4.0 License.

Open Access Earth System Science Data

Supplement of

Reconstructing ocean subsurface salinity at high resolution using a machine learning approach

Tian Tian et al.

Correspondence to: Lijing Cheng (chenglij@mail.iap.ac.cn)

The copyright of individual parts of the supplement might differ from the article licence.

S2.3.1 Comparison of typical machine learning reconstruction methods

Here we briefly introduce why feed-forward neural network (FFNN) is chosen in this study instead of other machine learning approaches.

(1) **Published researches.** Stamell et al. (2020) compared the advantages and disadvantages of three different machine learning methods in reconstructing the global ocean surface pCO2 from sparse observation data, and found that the extreme gradient boosting (XGBoost) produces the best pCO2 reconstruction overall, but the neural network (NN) method can be best generalized in poorly sampled regions and time periods. Wang et al. (2021) compared the performance of four machine learning algorithms XGBoost, the multi-linear regression (MLR), random forest (RF) and NN in estimating the subsurface temperature in the western Pacific, and proved that the neural network model outperformed the other three machine learning models. Lu et al. (2019) estimated subsurface temperature using cluster-neural network method, and showed that this method was superior to clustering linear regression and random forest method. The above-mentioned research shows that the NN method has superior generalization ability and is more robust for ocean data reconstruction.

(2) **Our experiments.** We have used "synthetic data" approach to test the performance of different machine-learning (ML). CNRM-CM6-1 high-resolution model simulation (Voldoire, 2019) is used (historical simulation that includes all climate forcings and is part of CMIP6). Because model results are with global coverage, dynamically consistent, thus can be used as "true" of salinity. We resample the model data according to the location of in situ observation to construct the "synthetic observations". The synthetic observations are prepared in two counterparts. The first is after further perturbation by observational errors/noises (denoted as "perturbed" data to account for the impact of observational errors). The observational errors are specificed in section 2.3.3. and Fig.2. The second is no perturbation (so the only error source of reconstruction is data sampling, this data is denoted as "non-perturbed"). Based on these synthetic data, we test different reconstruction schemes and compare the applicability of FFNN and light gradient boosting machine (LightGBM) (an improved method of XGBoost) to reconstruct globally ocean salinity data from sparse data. Fig. S1 and Fig.S2 show that the FFNN exhibits the lower RMSE (~0.035 psu) and the higher correlation coefficient (CC) (~0.866) compared with LightGBM for non-perturbed synthetic data. The perturbed data shows consistent results, although the addition of noises led to slightly higher RMSE and lower CC for both methods. In particular, perturbation of data induced a CC degradation of 3.5% and 6.4% for FFNN and LightGBM, respectively (Fig.S2d). Thus, the addition of observational noises leads to larger performance degradation for LightGBM than FFNN, suggesting that FFNN is more robust.

Finally, we also compare the performance of FFNN and LightGBM in reconstructing salinity using remote sensing and in situ observation data. The results show that the salinity field reconstructed by LightGBM had many non-continuous stripe-like structures (Fig.S3), which is apparently non-physical. This is associated with the intrinsic property of the LightGBM approach: 1) LightGBM discretize continuous variable into small bins by splitting the tree nodes; 2) Tree based models give stripe-like predictions as the model was trained using spatially sparse data.

With those tests and considerations, we found FFNN be an optimal choice to reconstruct subsurface salinity data in this study.

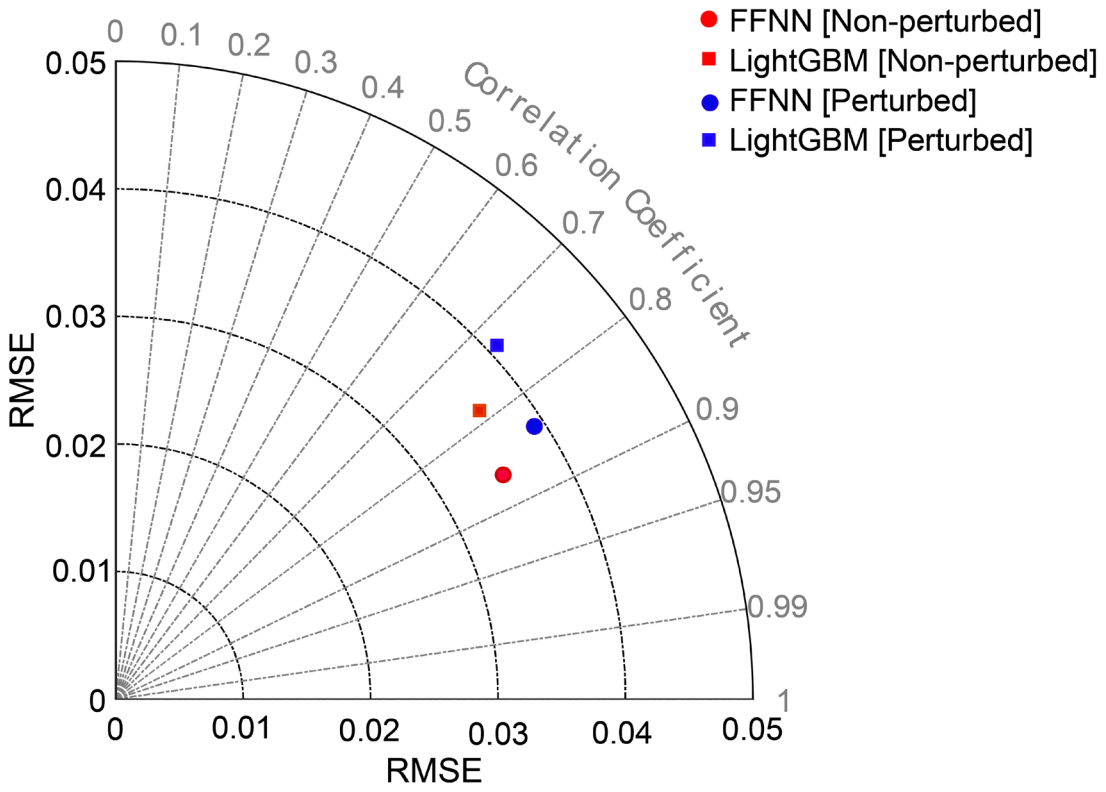

Fig.S1: Distribution of 1–2000 m averaged RMSE and correlation coefficient for different ML approaches. Blue markers represent the perturbed data; Red markers represent the non-perturbed data.

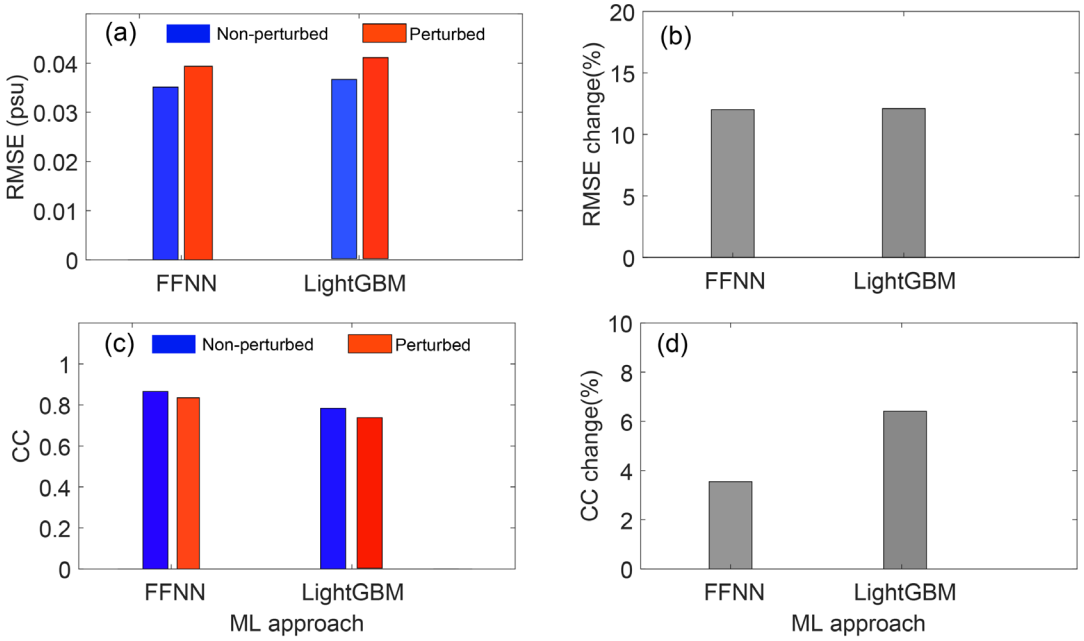

Fig.S2: Statistical metrics for FFNN and LightGBM methods. (a) 1–2000 m averaged RMSE. (b) Increase of RMSE from the non-perturbed data to the perturbed data; (c) 1–2000 m averaged correlation coefficient (CC). (d) Degradation of CC from the non-perturbed data to the perturbed data.

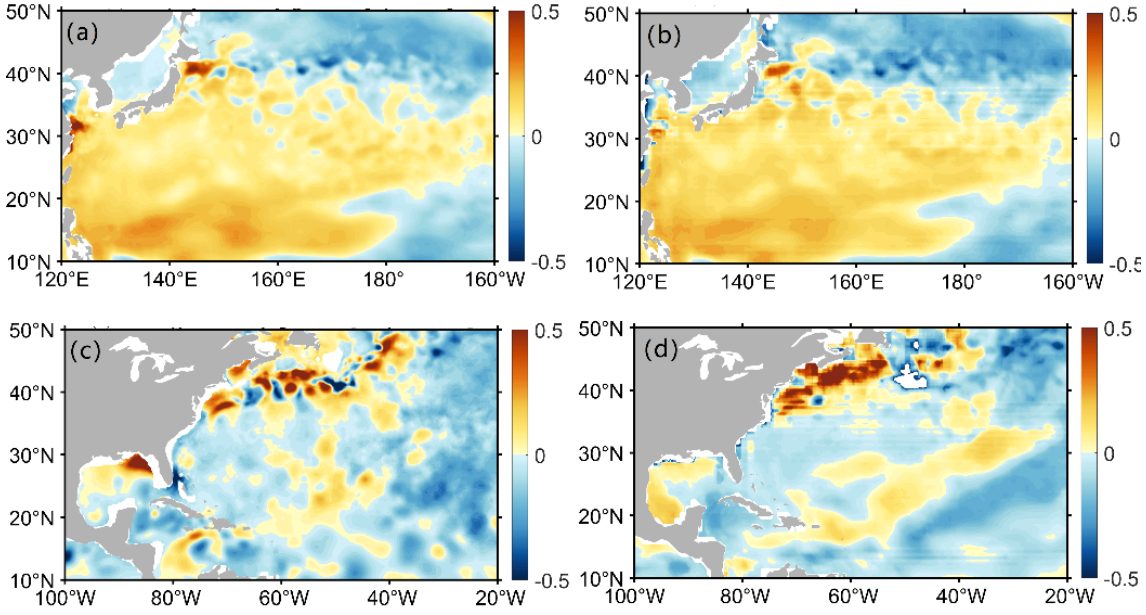

Fig.S3: The geographical distribution of salinity anomalies in the Kuroshio and Gulf Stream region from data reconstructed by FFNN (left) and LightGBM (right) method in January 2016: (a–b) Northwest Pacific region; (e–f) Northwest Atlantic region.

References

Gan, M., Pan, S., Chen, Y. ping, Cheng, C., Pan, H., and Zhu, X.: Application of the Machine Learning LightGBM Model to the Prediction of the Water Levels of the Lower Columbia River, J Mar Sci Eng, 9, 496, https://doi.org/10.3390/jmse9050496, 2021.

Lu, W., Su, H., Yang, X., and Yan, X. H.: Subsurface temperature estimation from remote sensing data using a clustering-neural network method, Remote Sens Environ, 229, 213–222, https://doi.org/10.1016/j.rse.2019.04.009, 2019.

Stamell, J., Rustagi, R., Gloege, L., and McKinley, G.: Strengths and weaknesses of three Machine Learning methods for pCO₂ interpolation, Geoscientific Model Development Discussions, 2020, 1–25, https://doi.org/10.5194/gmd-2020-311, 2020.

Voldoire, A.: CNRM-CERFACS CNRM-CM6-1-HR model output prepared for CMIP6 HighResMIP, Earth System Grid Federation, https://doi.org/10.22033/ESGF/CMIP6.1387, 2019.

Wang, H., Song, T., Zhu, S., Yang, S., and Feng, L.: Subsurface temperature estimation from sea surface data using neural network models in the western pacific ocean, Mathematics, 9, 852, https://doi.org/10.3390/math9080852, 2021.