



Fig. S6. Relative and absolute performance at radar stations withheld from the training dataset. Performance was best at interior sites, especially in the central and eastern United States. At a small minority (4%) of withheld sites, relative performance was poor ($R^2 < 0$, indicating performance was worse than a naive model that predicts the mean response for each observation), which may be due to local influences such as topography (e.g. see Florida Keys).