**Summary:**

Selecting an optimal AR model is challenging. In this study, the number of terms is determined based on three approaches: (1) inspection of auto-correlation function; (2) deterministic evaluation based on AIC/BIC. Three AR models are then built based on different results: AR13 (BIC), AR152 (auto-correlation), and AR378 (AIC). In order to further choose the possibly best one out of the three, the dataset is not only evaluated with training (2002-2017) and testing (2019-), but also split into validation period (2017-2018) for model selection.

It is found that AR378 could generate best results both in training period and validation period, and it is thus retained to perform forecast. The r2 for testing period is 0.79 and p value is 10-145 under null hypothesis, which indicates the correlation between observation and forecast is not trivial.

For forecast, four-week subset is chosen in Fig.5, but they cannot capture the trend very well, meaning this forecast model still performs poorly in the short-scale period. To further improve the model, two strategies are compared, one with white noise added to the model (Fig.6), and the other one removed seasonality (Fig.7). It is proved that removing seasonality could bring some improvements to the forecast skill while adding white noise does not improve a bit.

Given this context and the possibly best model, it can capture the seasonal variation of the temperature but fail to predict the fine-scale variations especially in short-period (daily scale). Thus, it possibly can be used in large scale forecast (long-term climatological forecast) but is challenging to well predict extreme situations or El Nino years. The sources of error are variant. From the model side, the linear combination of all past possible days is not able to describe the complex and dynamic system. Furthermore, one limitation is the data. Without enough data to fit in an optimal model, it is hard to reasonably generate good results.