



Towards a machine learning model for data assimilation and forecasting directly trained from observations

Christian Lessig, Peter Lean, Tony McNally, Mihai Alexe, Simon Lang, Matthew Chantry, and Peter Dueben

European Centre for Medium-Range Weather Forecasts (christian.lessig@ecmwf.int)

State-of-the-art data assimilation systems, such as the 4DVar system of the European Centre for Medium-Range Weather Forecasts (ECMWF), are highly successful in producing state estimates of the atmosphere constrained by millions of observations. However, existing systems require substantial approximations, e.g. in forward operators, and employ conventional models in their optimization loop. This limits the amount of information that can be extracted from observations, e.g. the assimilation of visible satellite channels is still challenging. The current approach also separates the use of observations into a data assimilation step and a forecasting one, with observations only being used indirectly for forecasting, for example for tuning of parametrizations and for evaluation.

Here, we explore the possibility to train large machine learning models for data assimilation and forecasting directly from observations. In particular, we build a generative transformer neural network that models the joint the probability distribution $p(y,x)$ over output states y for an input x . The input are observations from a temporal window, e.g. 6h or 12h, and the output y can either be an estimate of the state within the window or a short-term forecast, e.g. for another 12h. Different observations are processed by different embedding networks but then fused in the backbone transformer network. To obtain an integrated and consistent representation of the atmospheric state that corresponds to the different input data streams, we train with a variation of the masked token model training objective from natural language processing that impels the network to learn the correlation between the different input streams and channels. To properly represent the statistical nature of the estimation of y given x , our network provides an ensemble prediction as a nonparametric model for the probability distribution over y .

We present results for a network trained with a substantial amount of data, including different satellite observations (such as AMSU-A microwave sounders from NOAA 15-19 and the METOP satellites as well as IASI), radiosondes, and ground station-based measurements. The skill for both data assimilation and forecasting is analyzed and compared to ECMWF's operational 4DVar system. We also ablate the effect different observations have on the skill of the network output.