How to deal w__ missing input data

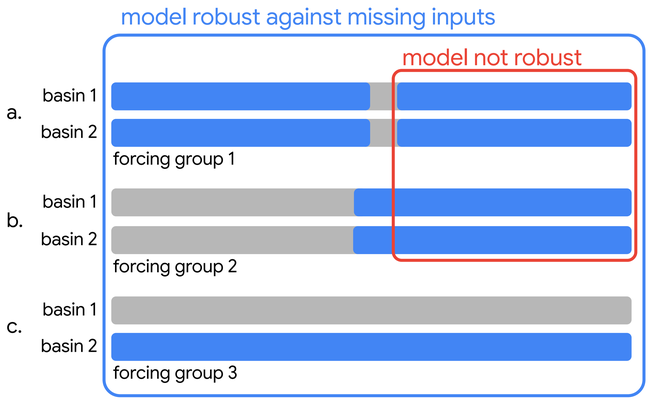
Martin Gauch, Frederik Kratzert, Daniel Klotz,
Grey Nearing, Deborah Cohen, Oren Gilon
gauch@google.com

Introduction

Deep Learning models are increasingly ubiquitous in the Earth sciences, both in research and applications.

The data needed to train and run these models is often messy, e.g.:
a. Randomly missing time steps (outages)
b. Data products with different temporal coverage
c. Data products with different spatial coverage

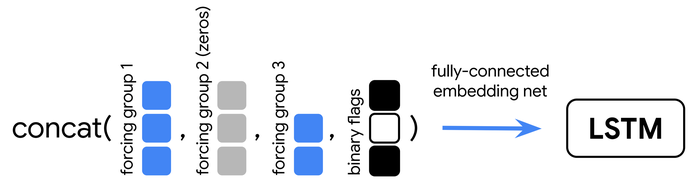
We propose different ways to build models that can deal with these discontinuities.



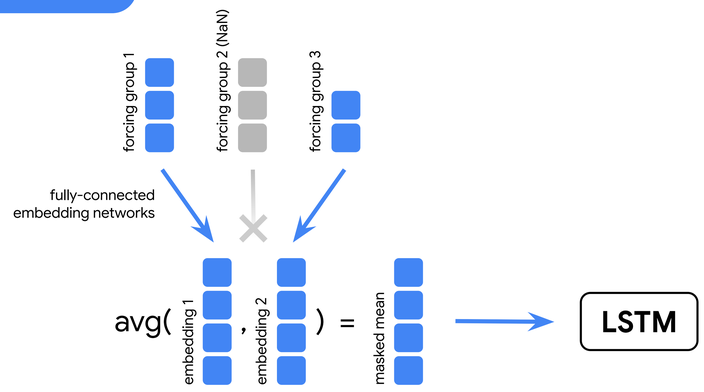
Methods

- **Input Replacing:**
  Replace missing inputs with a constant value (e.g., zeros) and add a binary flag to indicate these time steps.
- **Masked Mean Embedding:**
  Embed each group of input features to a shared embedding space, average non-NaN embeddings.
- **Attention:**
  Weighted averaging of embeddings based on static attributes.

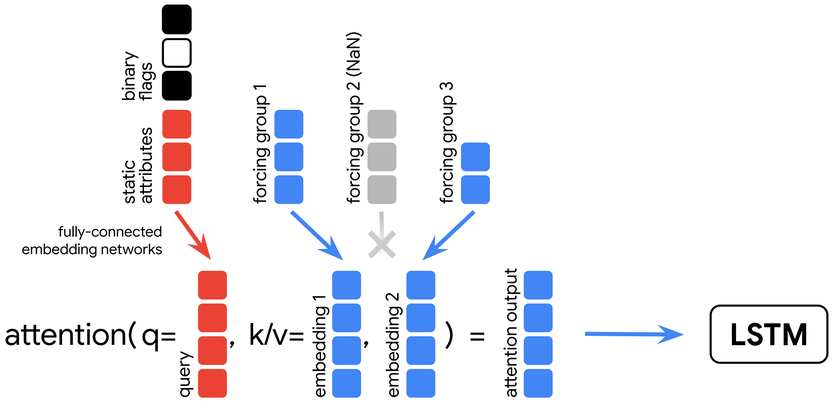
Input Replacing



Masked Mean



Attention

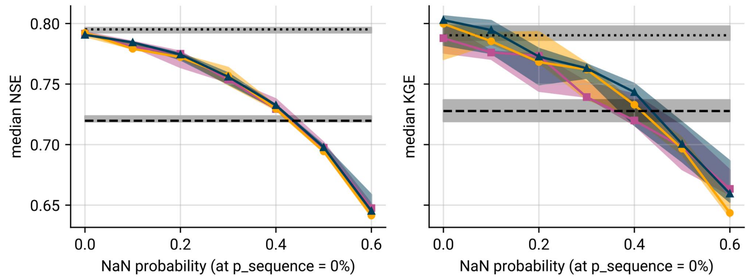


Results

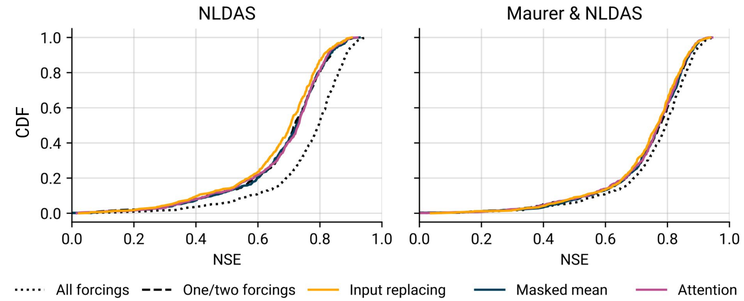
All modeling strategies can make good predictions even when certain input features are missing, but training must already reflect these conditions (e.g., by randomly dropping inputs during training).

Overall, Masked Mean Embedding seems the most promising:
it is simple and gives the best results.

We use the Masked Mean mechanism operationally in the models that run on g.co/floodhub.



Performance with increasing amount of missing input time steps



Performance when certain input products (NLDAS or Maurer and NLDAS) are missing entirely

Full paper:
hess.copernicus.org/articles/29/6221/2025

Google Research