

Are more data always better?

Machine learning forecasting of algal dynamics based on long-term observations from Blelham Tarn

Daniel Atton Beckmann

Mortimer Werther, Eleanor Mackay,
Evangelos Spyarakos, Peter Hunter, Ian Jones



Scottish Funding Council
Comhairle Maoineachaidh na h-Alba

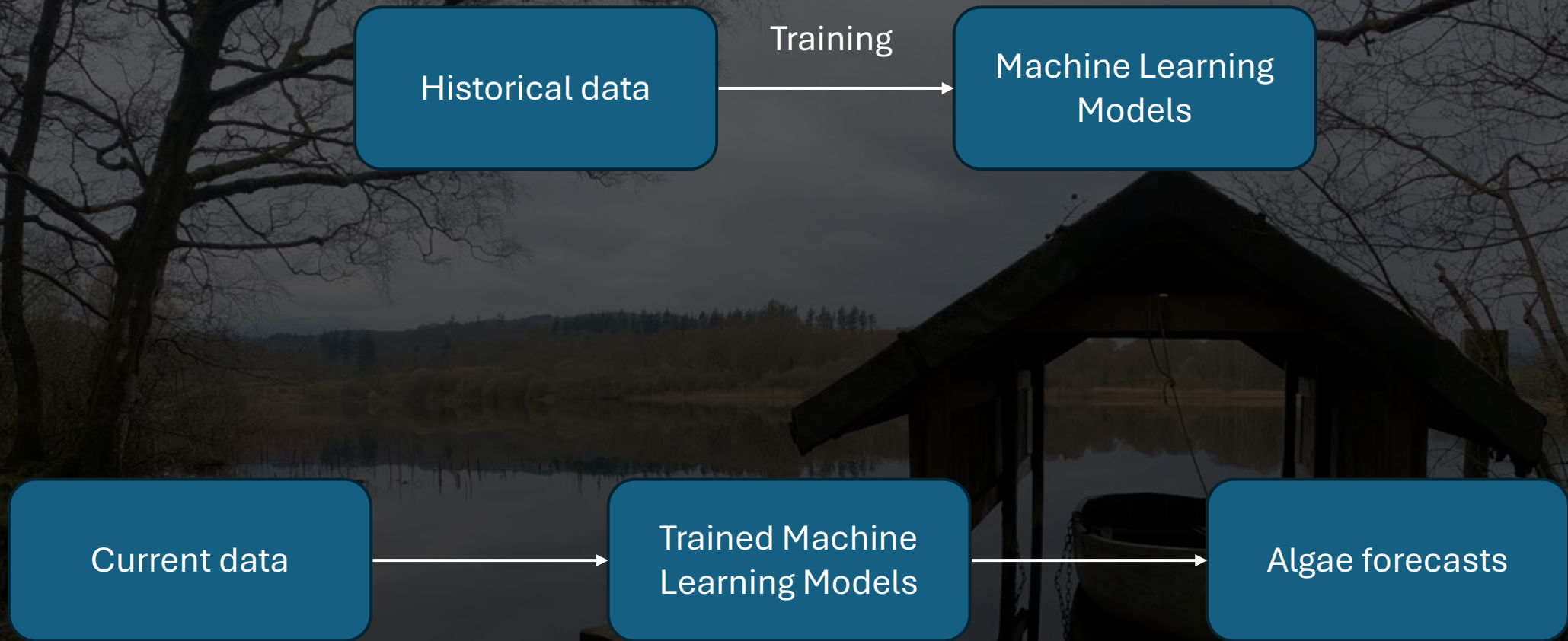
UNIVERSITY of
STIRLING



Scottish
Government
gov.scot

Hydro Nation Scholars Programme

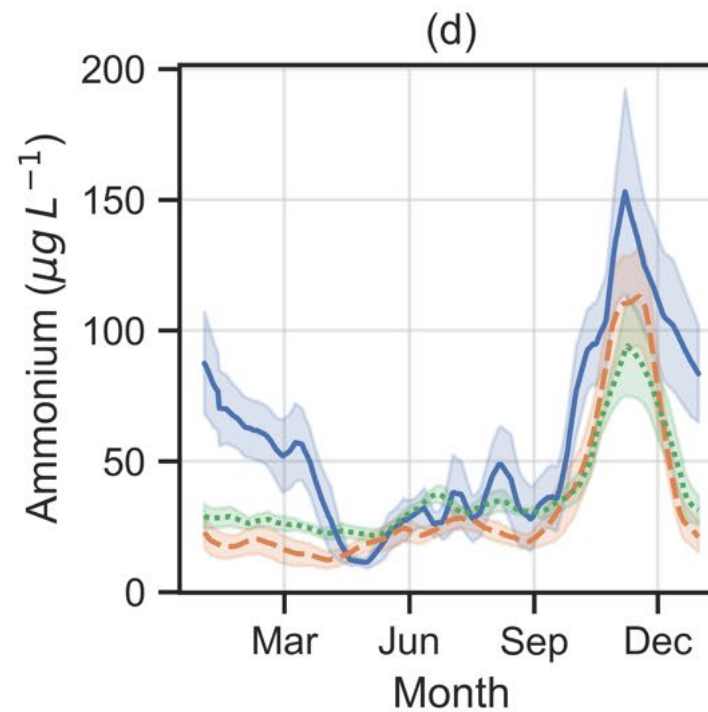
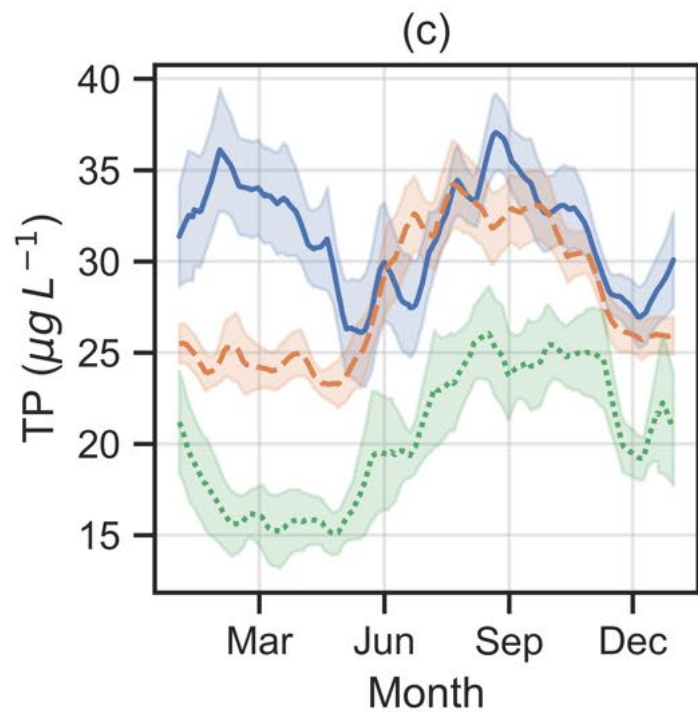
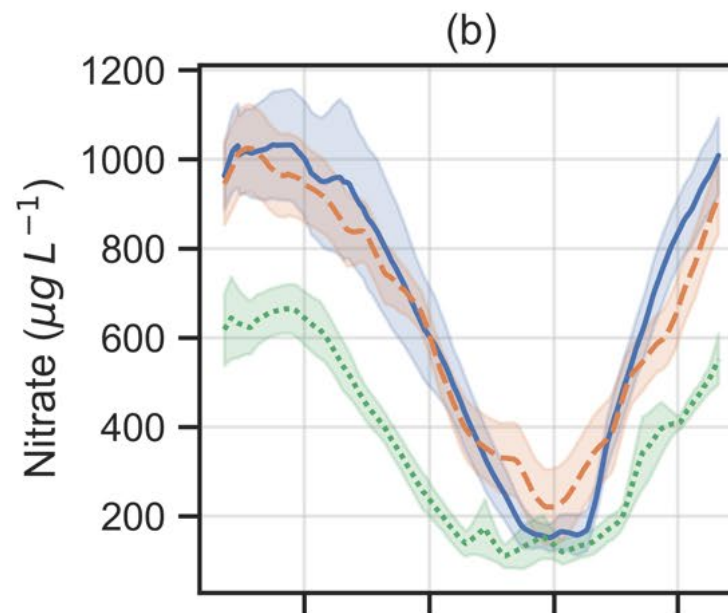
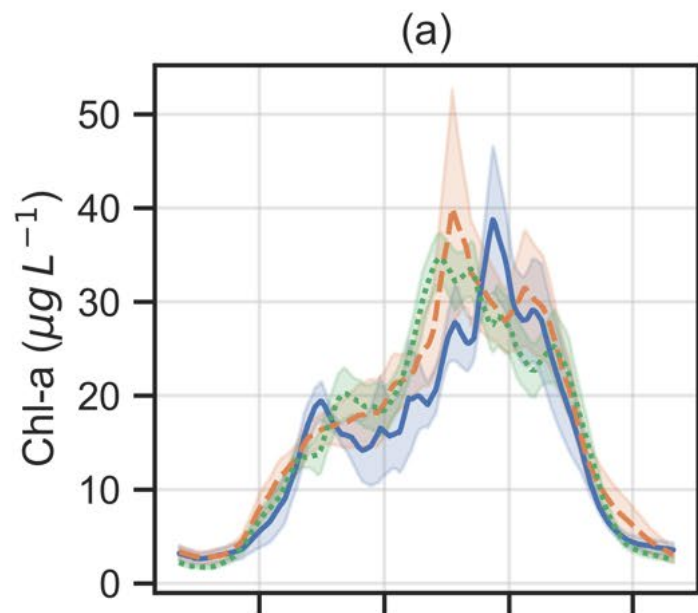




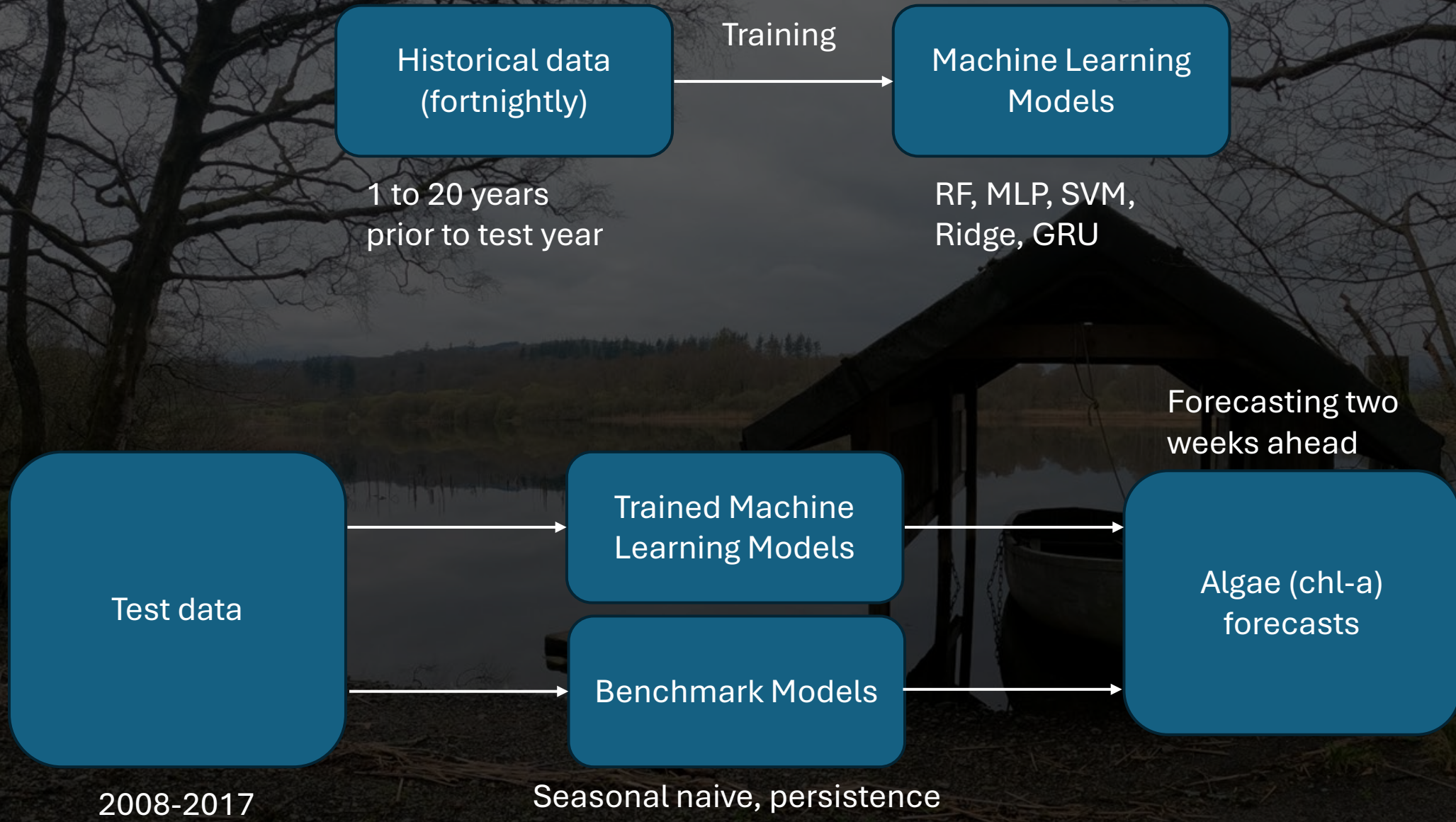
How much training data do we need?

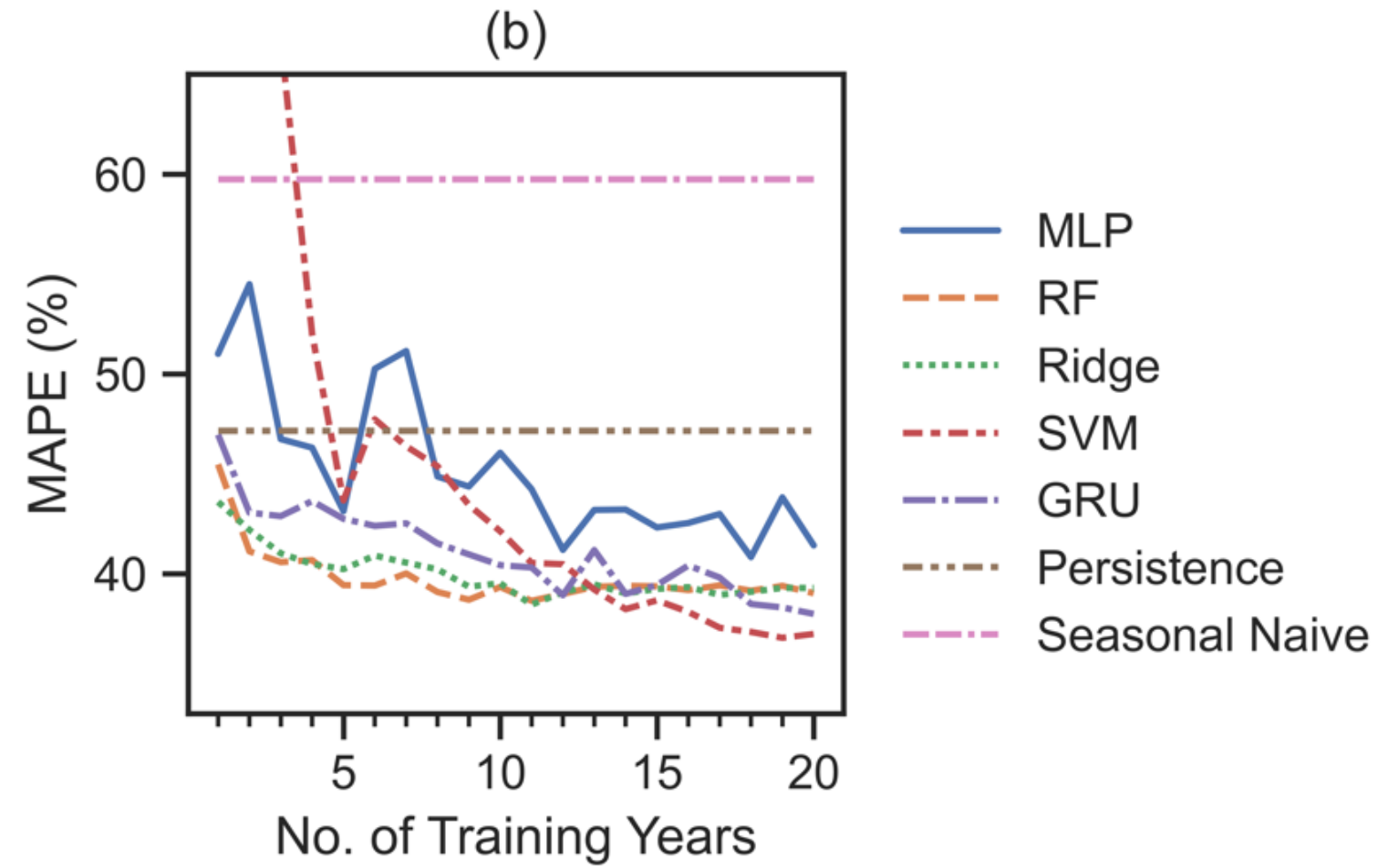
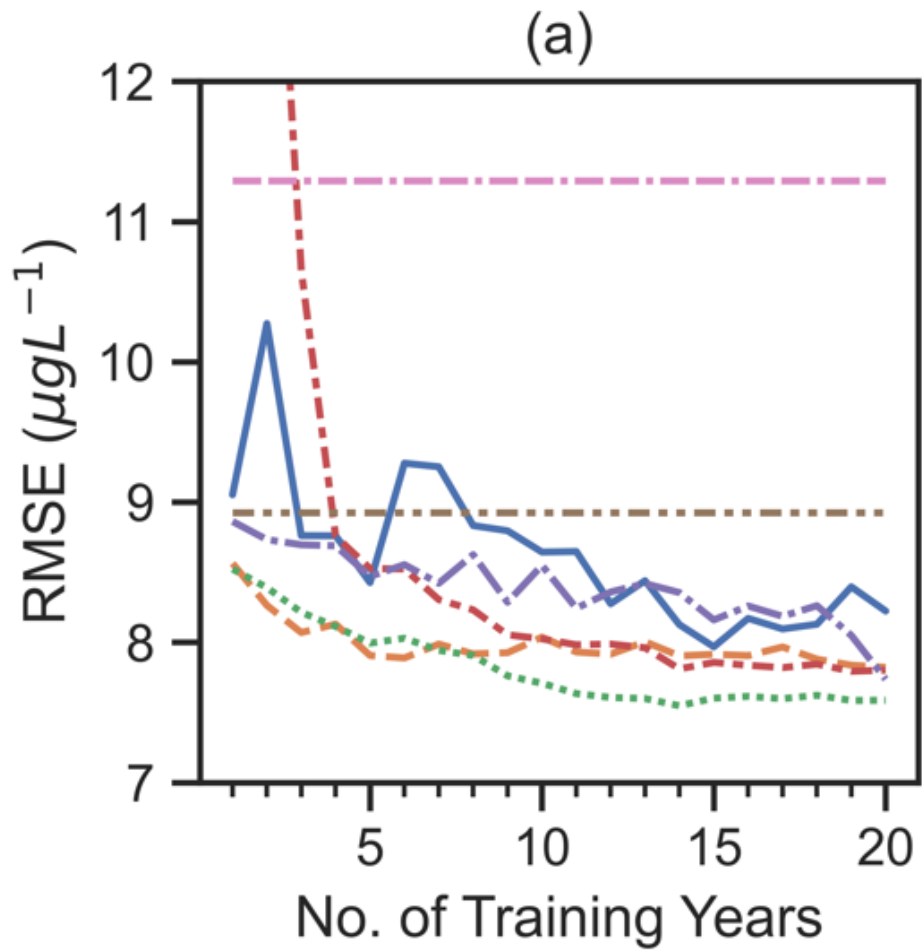


As much as possible?



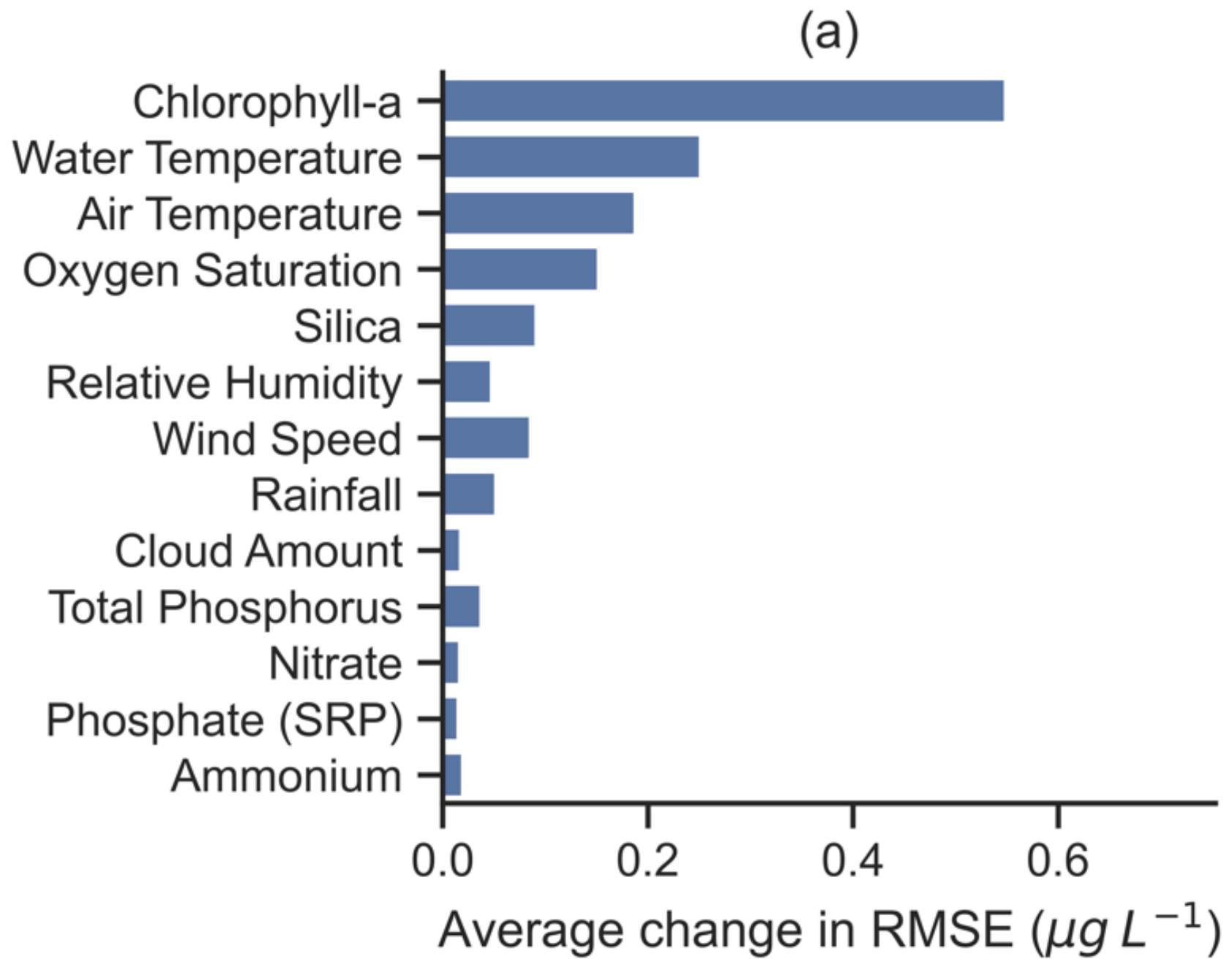
- 1987-1996
- - 1997-2007
- ... 2008-2017





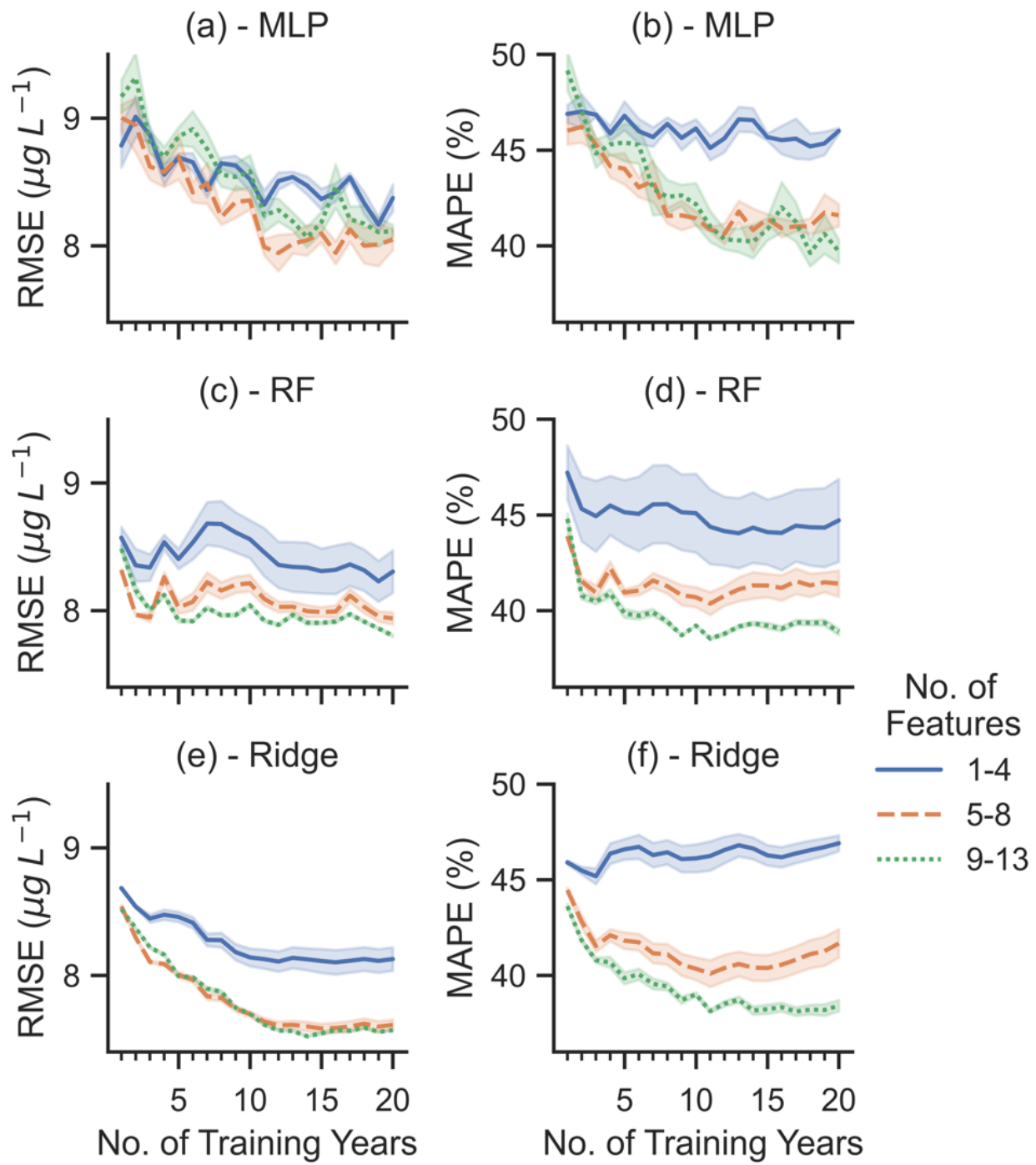
How does forecast performance change with the duration of training data?

Which variables are important?



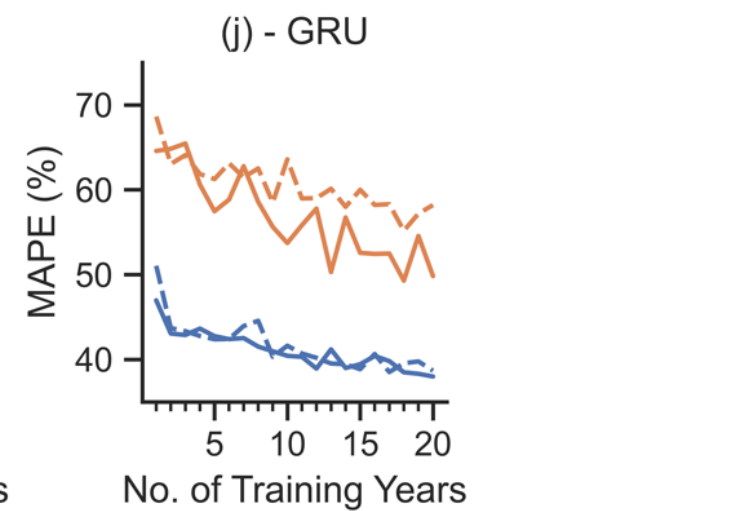
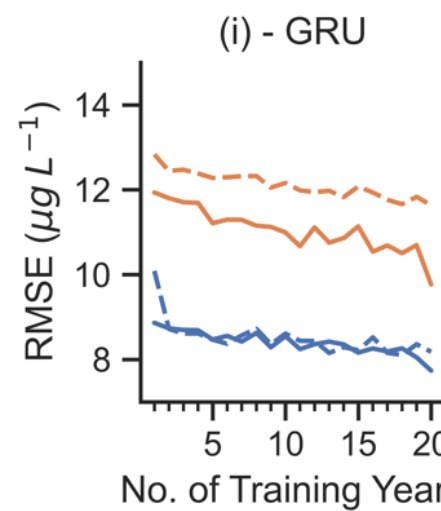
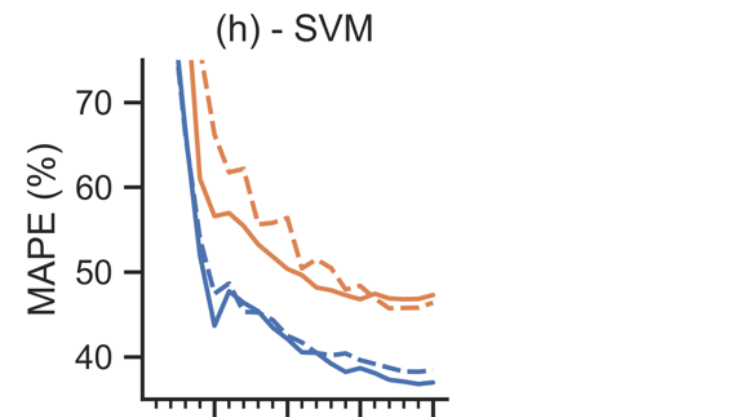
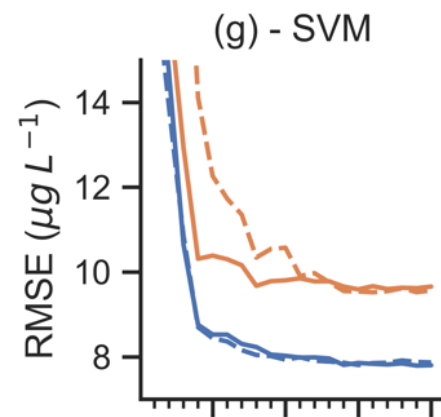
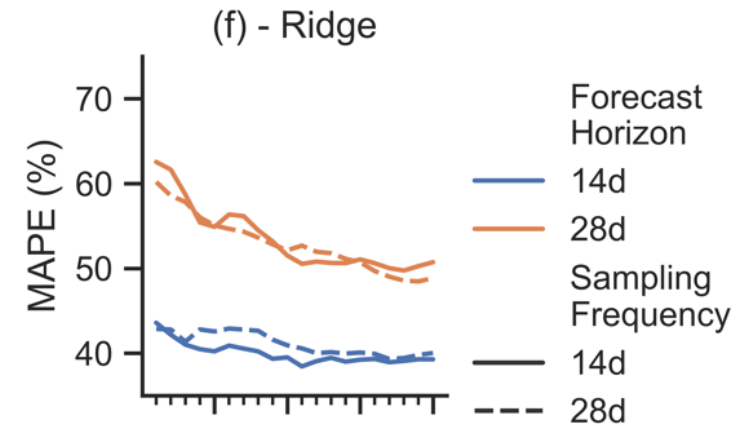
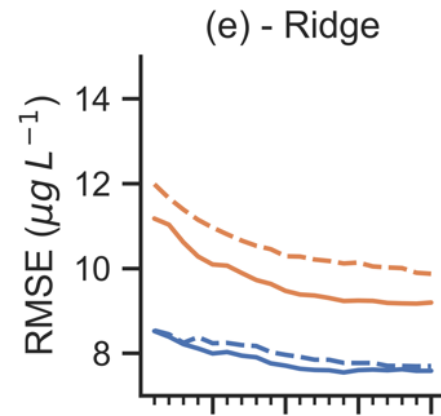


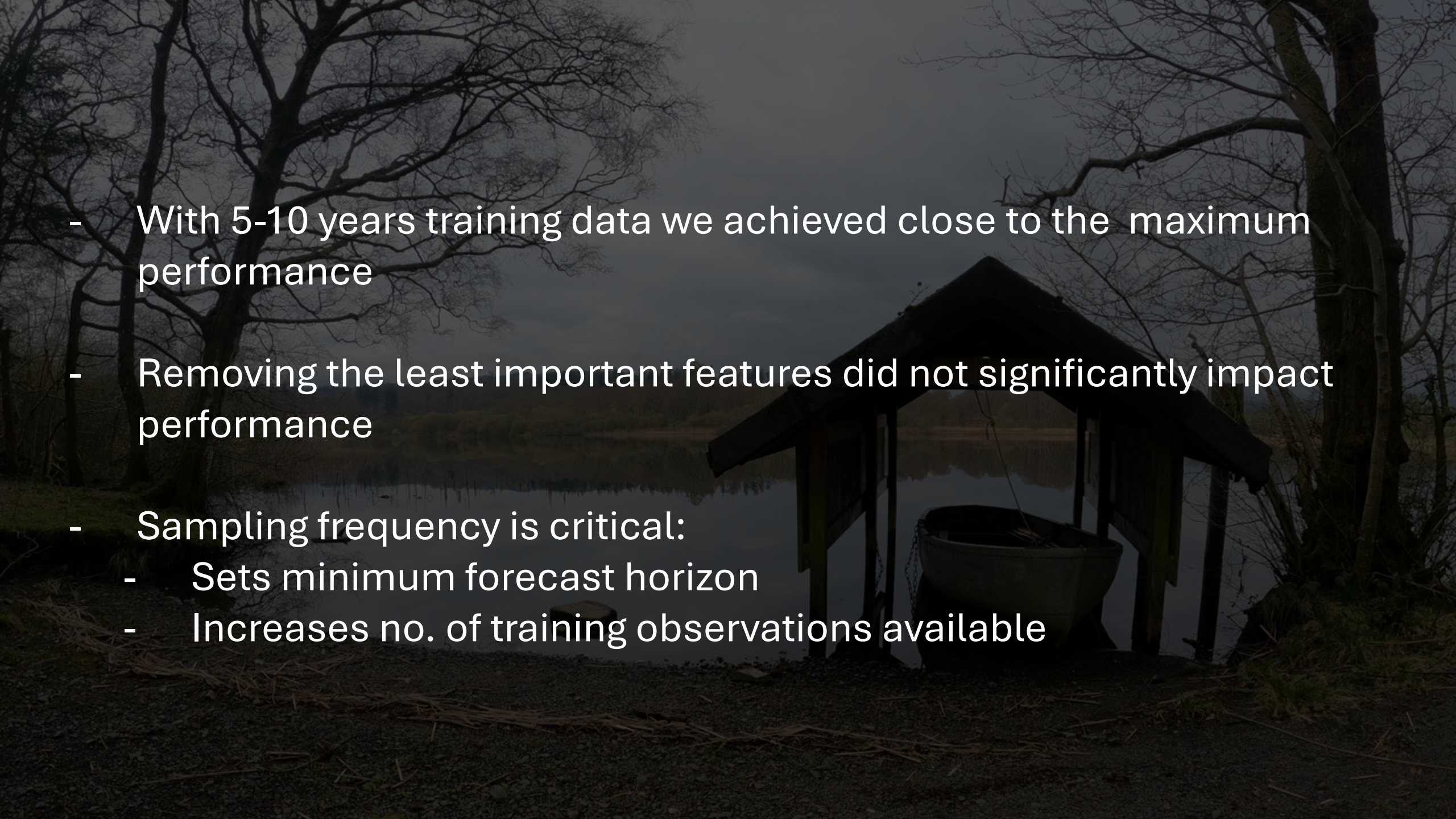
What happens if we remove less important variables?



What happens if we:

- forecast further into the future?
- Reduce the sampling frequency?



- 
- With 5-10 years training data we achieved close to the maximum performance
 - Removing the least important features did not significantly impact performance
 - Sampling frequency is critical:
 - Sets minimum forecast horizon
 - Increases no. of training observations available

