

# Getting ready for the workshop

Workshop materials:

<https://github.com/OlssonF/Forecast-evaluation-EFI25>

If you want to code along:

1. Download/fork/clone code
2. Install the packages

---

# A practical starter to Forecast Evaluation and Synthesis

---

Ecological Forecasting Initiative  
May 2025

---

Freya Olsson, Caleb Robbins, Quinn Thomas

---

# Overview and Goals

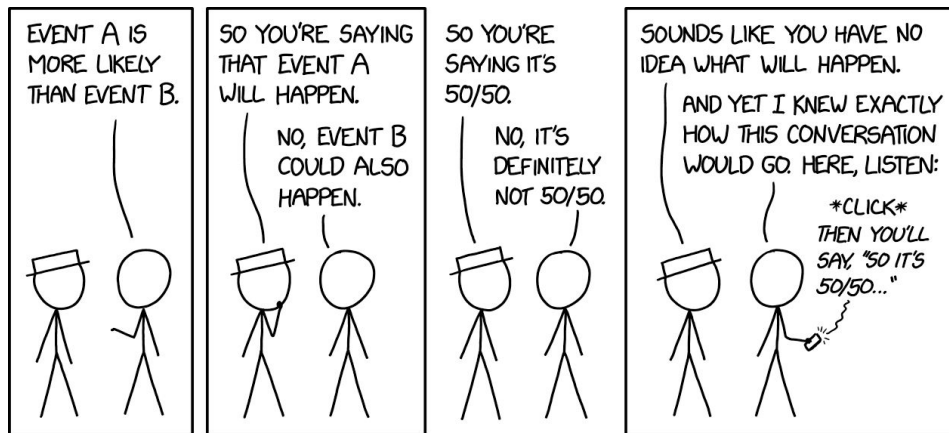
## Introduction

- Forecast evaluation
- Evaluation metrics
- Forecast catalogues

## R code demonstration

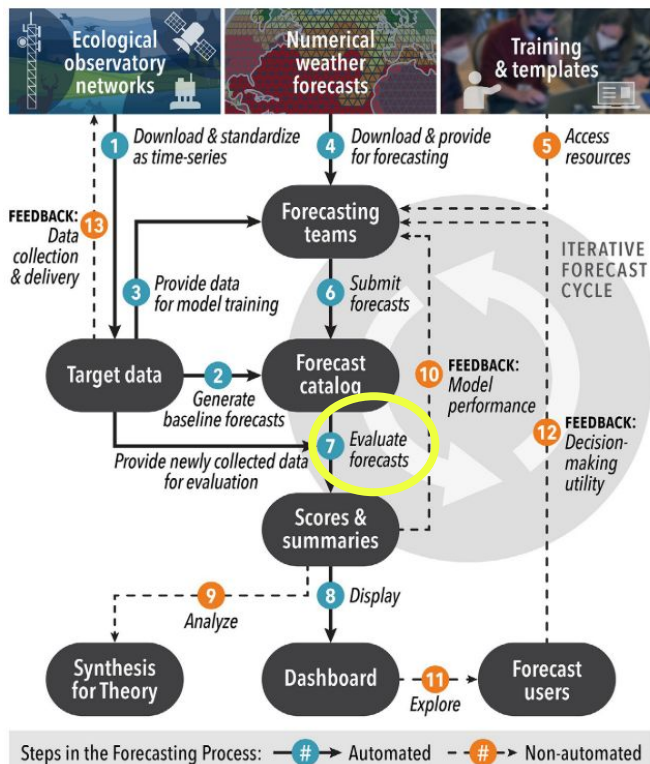
- Visualising evaluation metrics
- Summarising performance
- Comparing forecast performance

So, you made a forecast of future ecological conditions - now what?

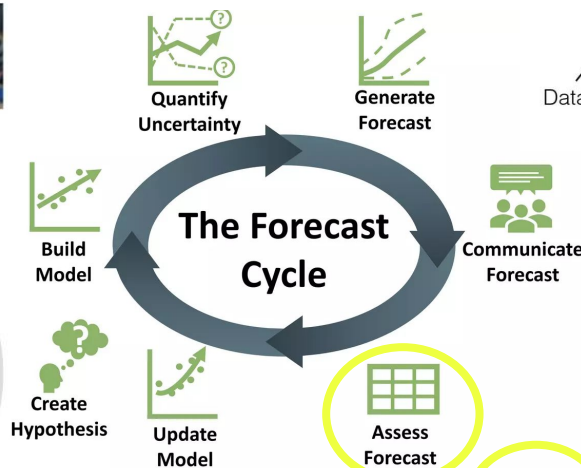


<https://xkcd.com/2370>

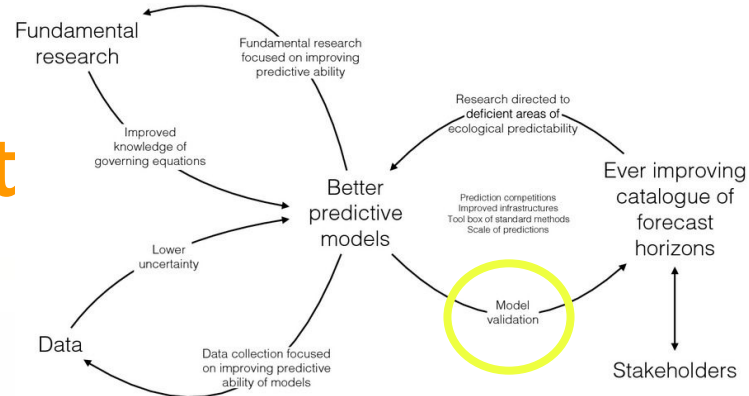
# Evaluation within the forecast



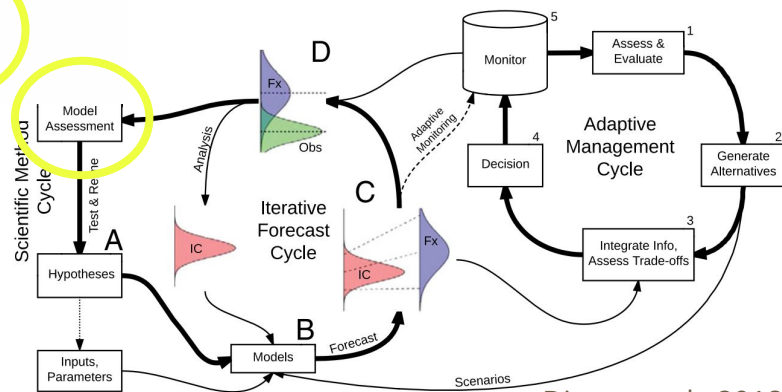
Thomas et al., 2024



Moore et al., 2022



Petchey et al., 2015



Dietze et al., 2018

# Using forecast evaluation to inform model development and synthesise knowledge

## What can we learn from the error?

Bias, spread, probability place on observation,  
confidence interval reliability

Iterative forecast model  
improvement - the  
forecasting cycle!

Patterns of forecast  
performance across:

- Space
- Time
- Model

# Fit-for-purpose forecast evaluation

1. Dependent on context and use
2. Many conversations on forecast evaluation from philosophical to statistical
  - a. Lots to be learned from other fields!
3. Our experiences from a large scale synthesis effort as part of the NEON Forecasting Challenge

Jacobs, B., Tobi, H., & Hengeveld, G. M. (2024). Linking error measures to model questions. *Ecological Modelling*, 487(July 2023), 110562.

<https://doi.org/10.1016/j.ecolmodel.2023.110562>

Parker, W. S. (2020). Model Evaluation: An Adequacy-for-Purpose View. *Philosophy of Science*, 87(3), 457–477. <https://doi.org/10.1086/708691>

Smith, L. A., Suckling, E. B., Thompson, E. L., Maynard, T., & Du, H. (2015). Towards improving the framework for probabilistic forecast evaluation.

*Climatic Change*, 132(1), 31–45. <https://doi.org/10.1007/s10584-015-1430-2>

Simonis, J. L., White, E. P., & Ernest, S. K. M. (2021). Evaluating probabilistic ecological forecasts. *Ecology*, 102(8), 1–8.

<https://doi.org/10.1002/ecy.3431>

# NEON Forecasting Challenge

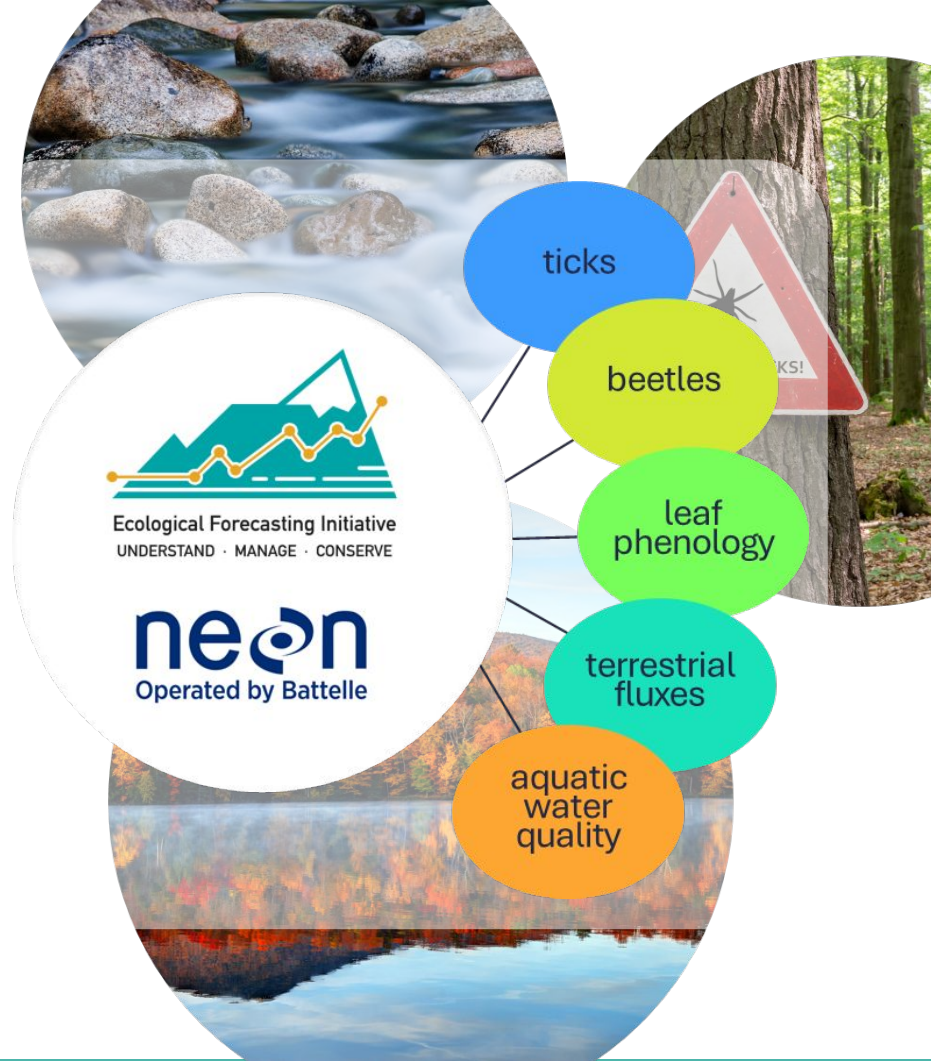
- 5 themes, 10 variables – communities, populations, phenology, ecosystem fluxes and states (terrestrial and aquatic)
- The NEON Challenge has > 5 million forecast-observation pairs!

## Learn more!

Thomas et al. (2023). The NEON Ecological Forecasting Challenge.

Front. Ecol. Environ., 21(3), 112–113.

[neon4cast.org](https://neon4cast.org)





# The catalog is open-source

- ★ Raw forecasts
- ★ Observations (targets)
- ★ Evaluated forecasts (scores)
  - Observation
  - Forecast stats
  - Performance metrics

We will use this in part 2 of the workshop!

## NEON Ecological Forecasting Challenge Catalog

Source Share Language: English

Up Browse

### Description

A STAC (Spatiotemporal Asset Catalog) describing forecasts and forecast scores for the neon4cast Forecasting Challenge

### Catalogs

Tiles List Ascending Descending

Filter catalogs by title, description or keywords



#### Forecast Summaries

**Parquet** Summaries are the forecasts statistics of the raw forecasts (i.e., mean, median, confidence intervals). You can access the summaries at the top level of the dataset where all...

2022-01-01 00:00:00 UTC - 2026-12-27 00:00:00 UTC



#### Inventory

**Parquet** The catalog contains forecasts for the NEON Ecological Forecasting Challenge. The forecasts are the raw forecasts that include all ensemble members (if a forecast...

2025-02-18 00:00:00 UTC - 2026-04-13 00:00:00 UTC



#### NOAA-Forecasts

**Parquet** The catalog contains NOAA forecasts used for the NEON Ecological Forecasting Challenge. The forecasts are the raw forecasts that include all ensemble members (if a forecast...

2020-01-01 00:00:00 UTC - 2025-04-20 00:00:00 UTC



#### Site Metadata

**Parquet** The catalog contains site metadata for the NEON Ecological Forecasting Challenge

2013-03-07 00:00:00 UTC - 2025-04-19 00:00:00 UTC



#### Forecasts

**Parquet** Forecasts are the raw forecasts that includes all ensemble members or distribution parameters. Due to the size of the raw forecasts, we recommend accessing the scores...

2017-02-01 00:00:00 UTC - 2025-04-17 00:00:00 UTC



#### Scores

**Parquet** The catalog contains scores for the NEON Ecological Forecasting Challenge. The scores are summaries of the forecasts (i.e., mean, median, confidence intervals), matched...

2017-02-01 00:00:00 UTC - 2025-04-17 00:00:00 UTC



#### Targets

**Parquet** The targets are observations that can be used to evaluate and build forecasts. We provide the code to access different targets as an asset.

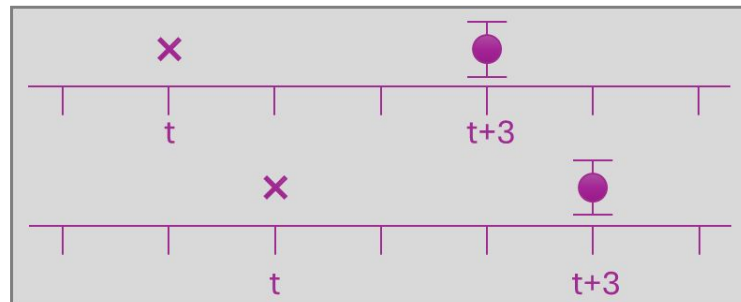
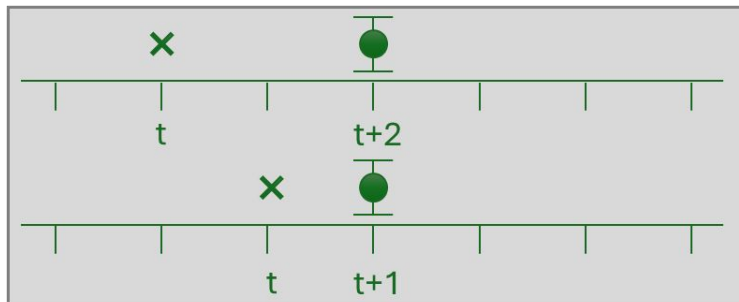
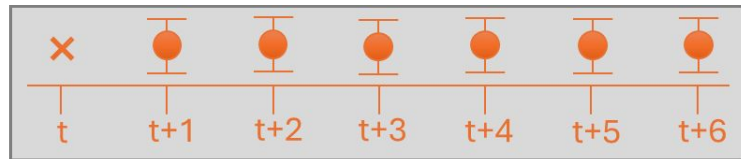
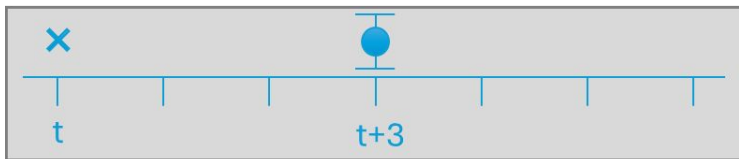
2013-03-07 00:00:00 UTC - 2025-04-19 00:00:00 UTC



# Some forecast definitions

A forecast is a prediction of ***future conditions with specified uncertainty***

Forecasts can be produced ***iteratively*** (multiple times) – this might be producing a forecast for the same time point multiple times or shifting the window of the forecast.



# Some forecast definitions

**Reference date, start date, or issuance date** = date of forecast generation. For a genuine forecast that would be today!

**Horizon, lead time** = time into the future of the forecast, difference between reference date and forecasted date!

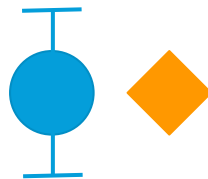
reference date  
datetime

A forecast made on January 1st for January 2nd would have a horizon of 1 day!

A diagram with two arrows pointing from the words 'reference date' and 'datetime' to the dates 'January 1st' and 'January 2nd' respectively in the sentence below. 'reference date' is in teal and 'datetime' is in teal. The arrows are thin black lines.

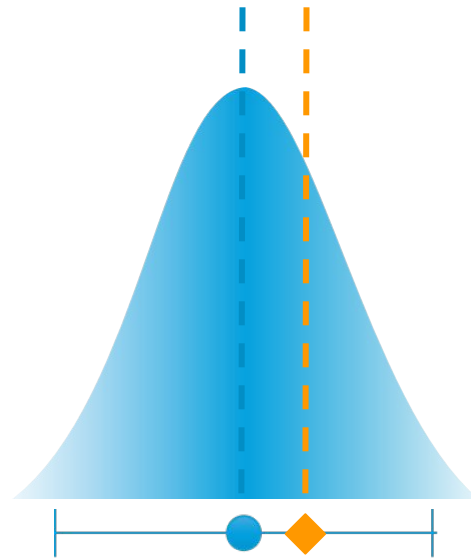
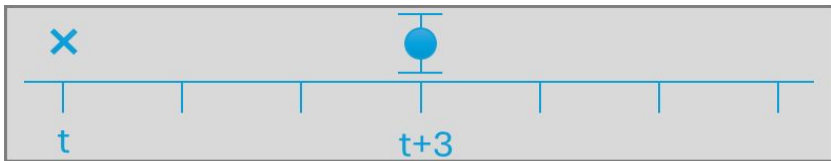
# Introduction to forecast evaluation

The basis to evaluation is the comparison of a **forecast**-**observation** pair



Can only occur once observations are available

Following this example, we could evaluate the forecast at  $t+3$  when 3 timesteps have passed



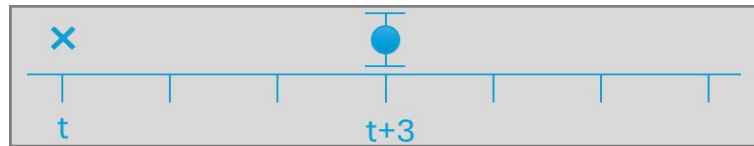
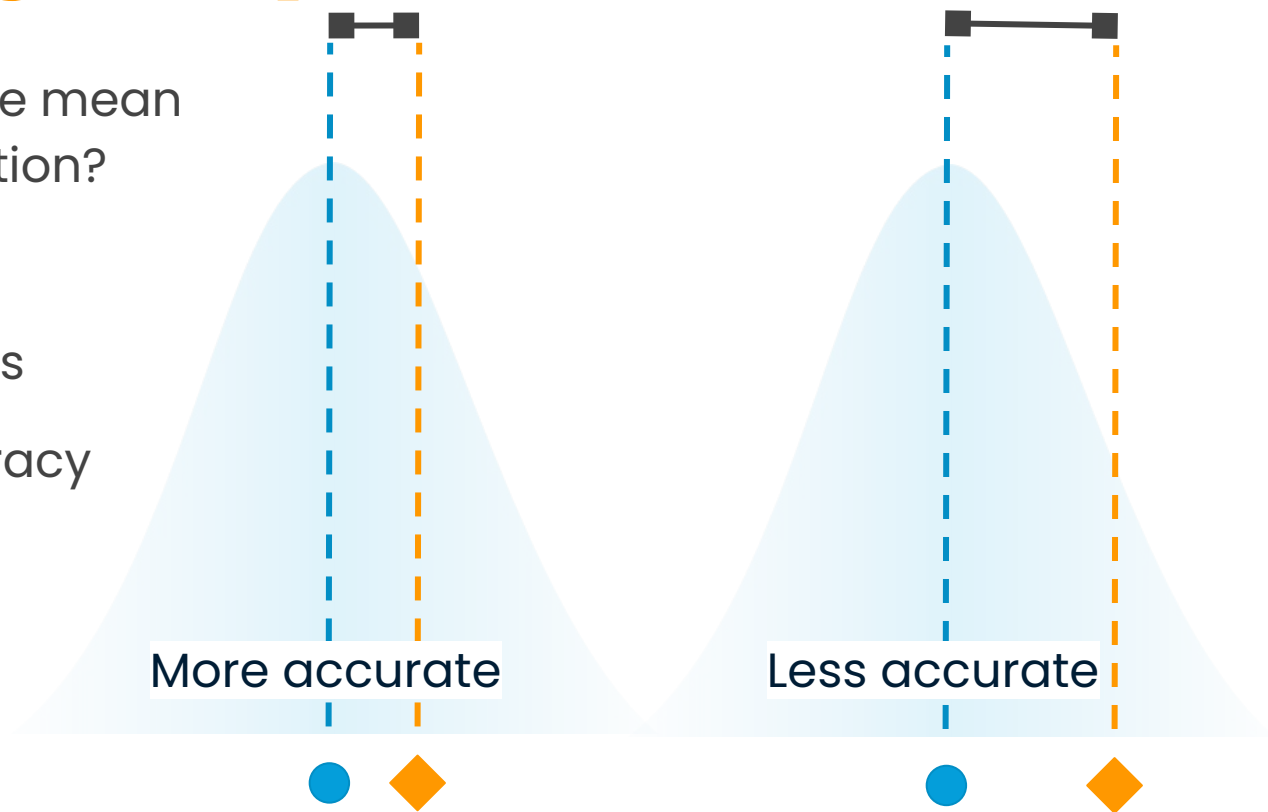
# Evaluating mean predictions

How close is the mean to the observation?

Error and bias

Multiple metrics

Evaluate accuracy

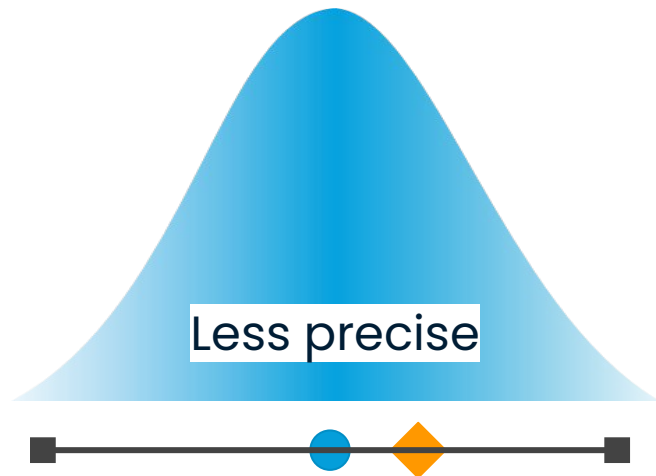
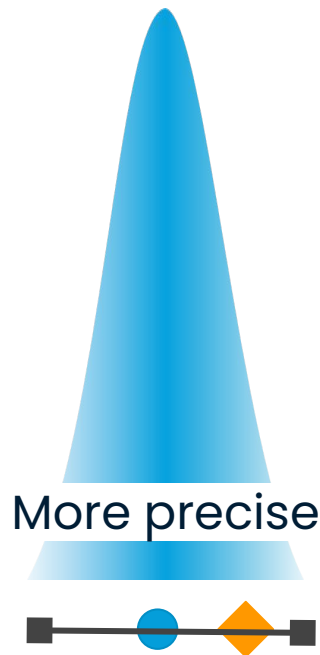
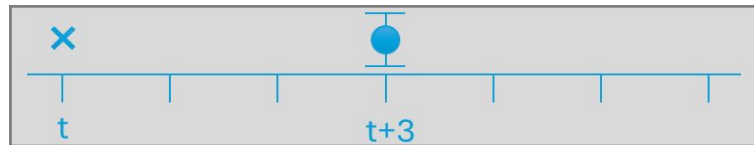


# Evaluating prediction spread

How wide is the distribution?

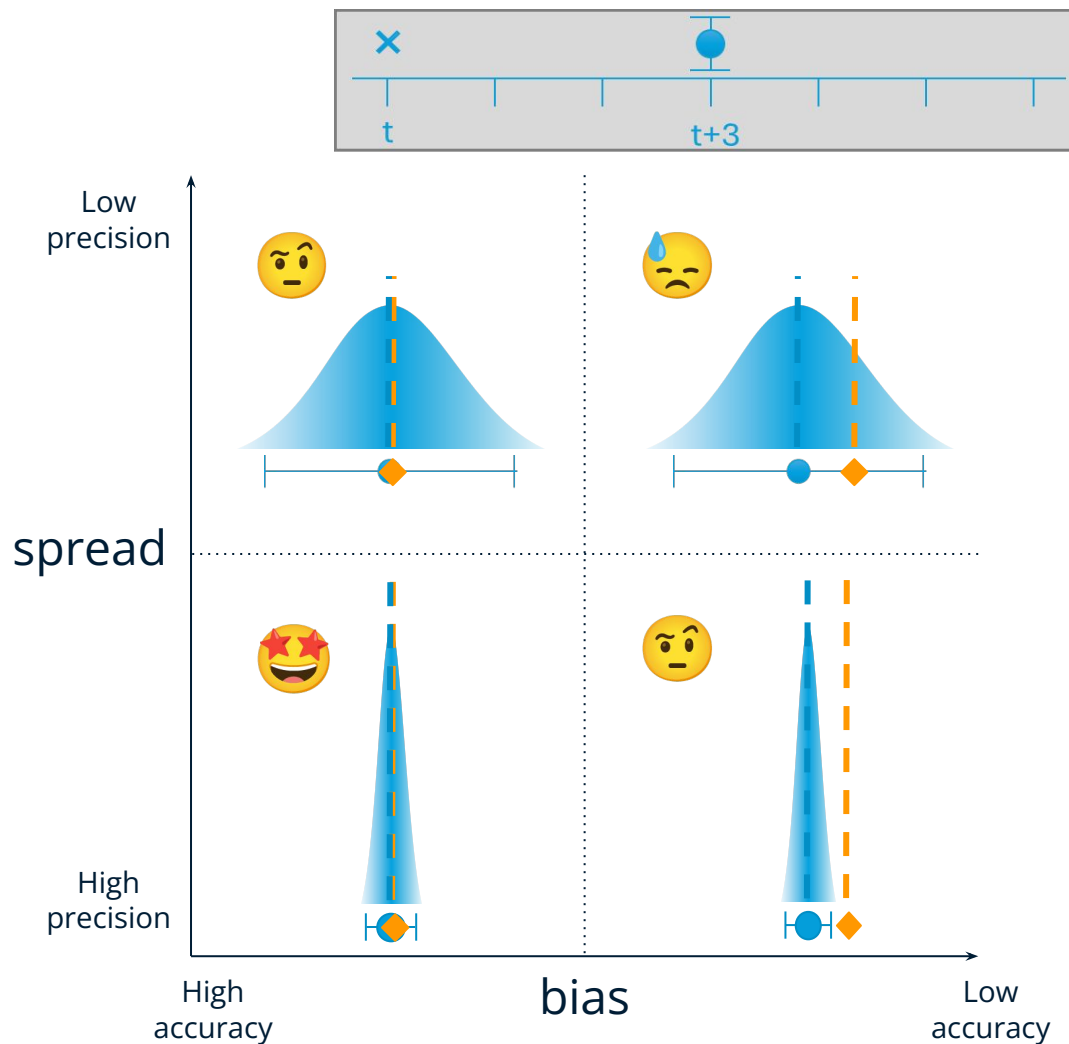
Quantified by standard deviation or other measures of spread

Evaluates forecast precision

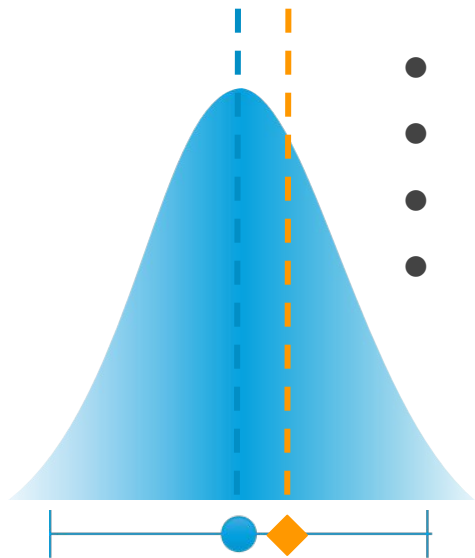


# Distributional or probabilistic forecast evaluation

Evaluating the **precision** and **accuracy** of the forecast



# Continuous rank probability score (CRPS)

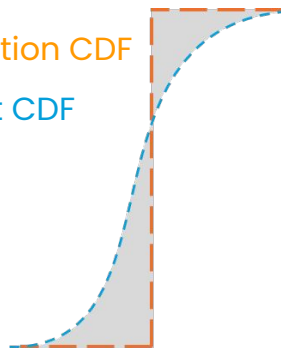


Convert the *probability density function* into a *cumulative distribution function*

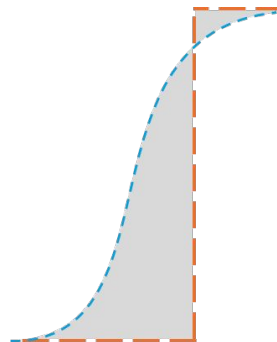
- Lower CRPS is better
- Same unit as the observation
- Considers the distribution of the forecasts as a whole
- Generalisation of mean absolute error for probabilistic predictions

Observation CDF

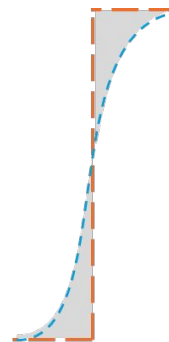
Forecast CDF



Change in  
accuracy



Change in  
precision

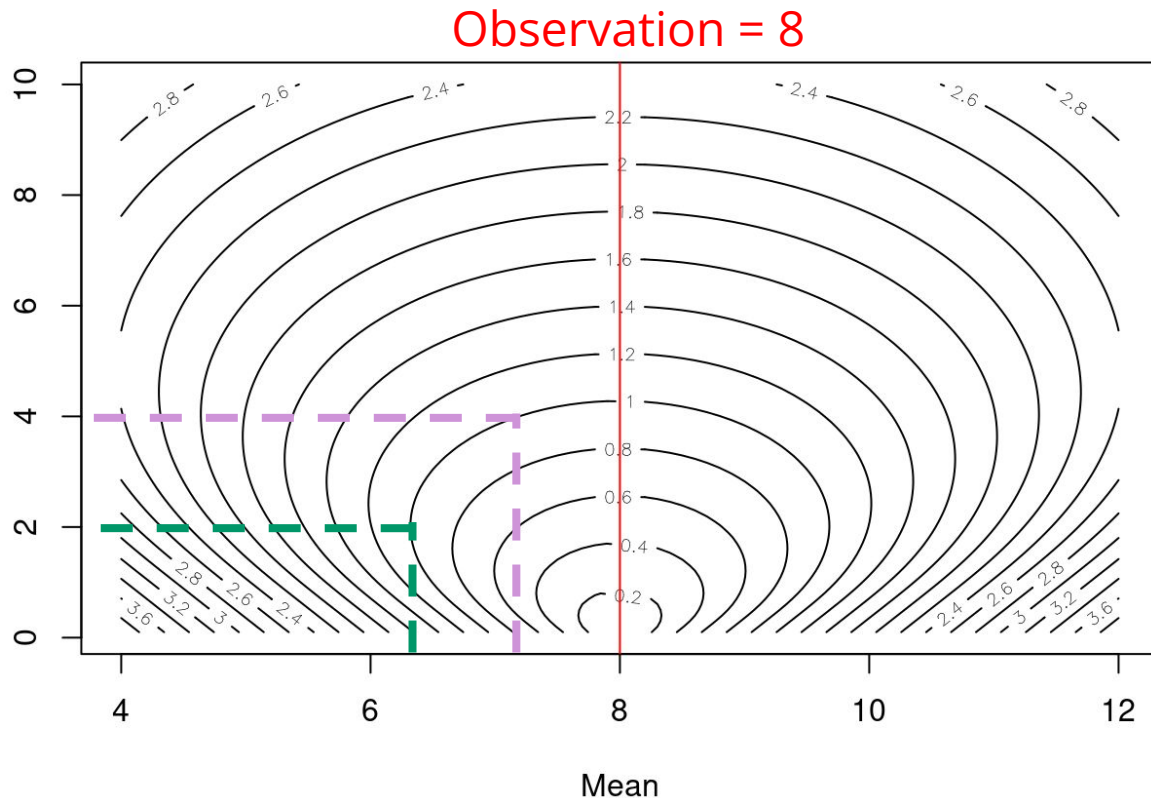




# Continuous rank probability score

Same CRPS when you are not quite as accurate but a bit more precise!

SD



# Other probabilistic evaluation metric

Also calculated as part of the NEON Challenge evaluation:

- Logarithmic (log) score = Probability that the forecast places on the observed outcome
- Forecast summaries (mean, median, 90 and 95% confidence intervals)

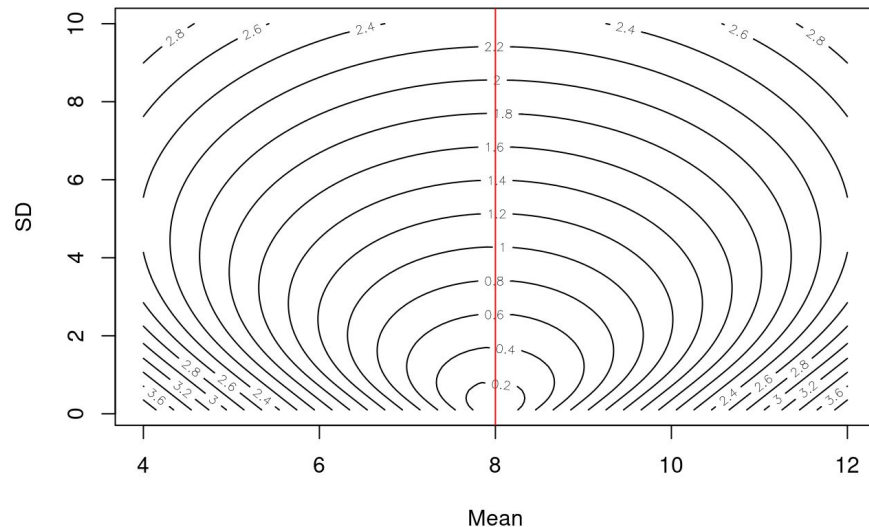
Simonis, J. L., White, E. P., & Ernest, S. K. M. (2021). Evaluating probabilistic ecological forecasts. *Ecology*, 102(8), 1–8. <https://doi.org/10.1002/ecy.3431>

Jolliffe, I. T., & Stephenson, D. B. (2012). *Forecast verification: a practitioner's guide in atmospheric science*. (Ian T. Jolliffe & D. B. Stephenson, Eds.) (2nd ed.). Oxford: Wiley Blackwell.

Gneiting, T., & Raftery, A. E. (2007). Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association*, 102(477), 359–378. <https://doi.org/10.1198/016214506000001437>

# Calibration of the confidence intervals important for forecast performance

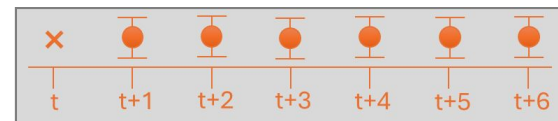
Seen that **forecasts do when the spread is low** but there are also conditions under which low spread could be **bad for forecasts – when they are inaccurate**



# Reliability of the confidence intervals

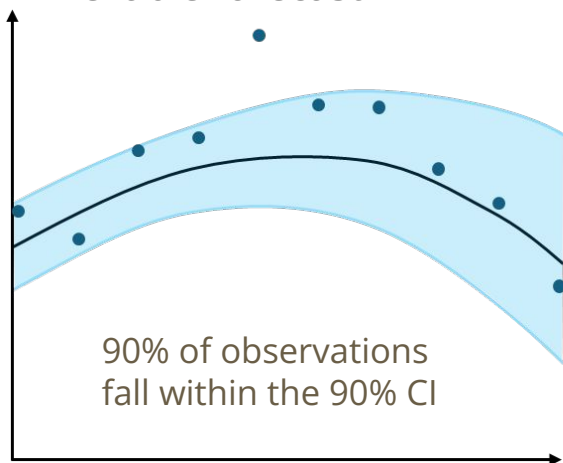
Reliability of confidence or predictive intervals

Also sometimes referred to as **calibration**

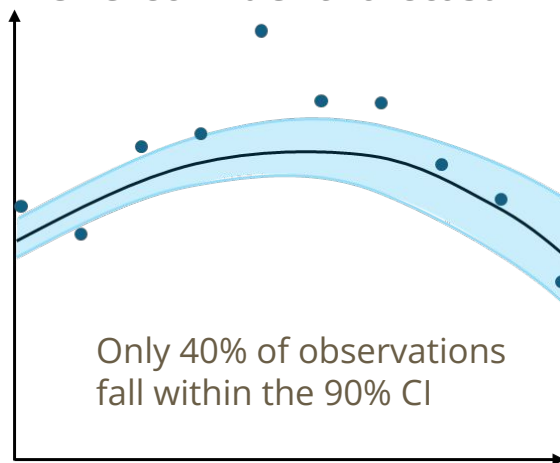


Plots show a forecast with 90% predictive intervals (plus observation points).

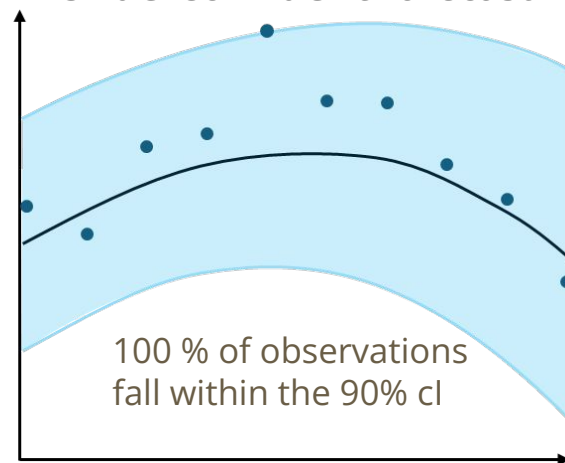
Well calibrated CI,  
Reliable forecast



Poorly calibrated CI,  
**Overconfident** forecast



Poorly calibrated CI,  
**Underconfident** forecast



# Using forecast evaluation to inform model development and synthesise knowledge

## What can we learn from the error?

Bias, spread, probability place on observation,  
confidence interval reliability

Iterative forecast model  
improvement - the  
forecasting cycle!

Patterns of forecast  
performance across:

- Space
- Time
- Model

# Not enough to just generate a single forecast

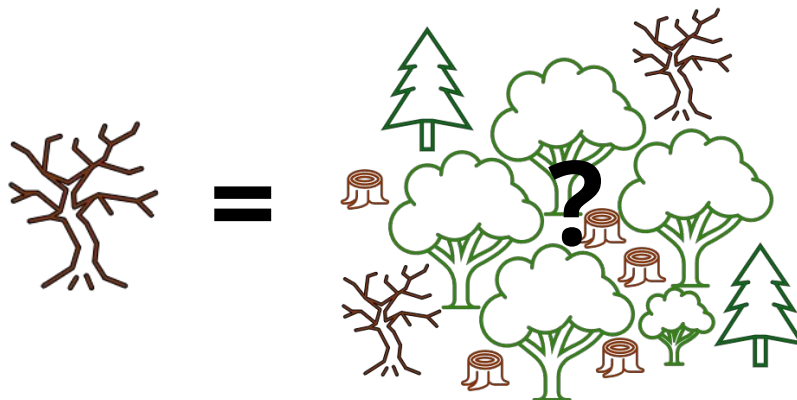
Why?

- Evaluating multiple forecasts
  - patterns in performance
- Comparing with forecasts generated by other models

How?

- Existing forecast catalogues
- Generating multiple forecasts - building a catalogue

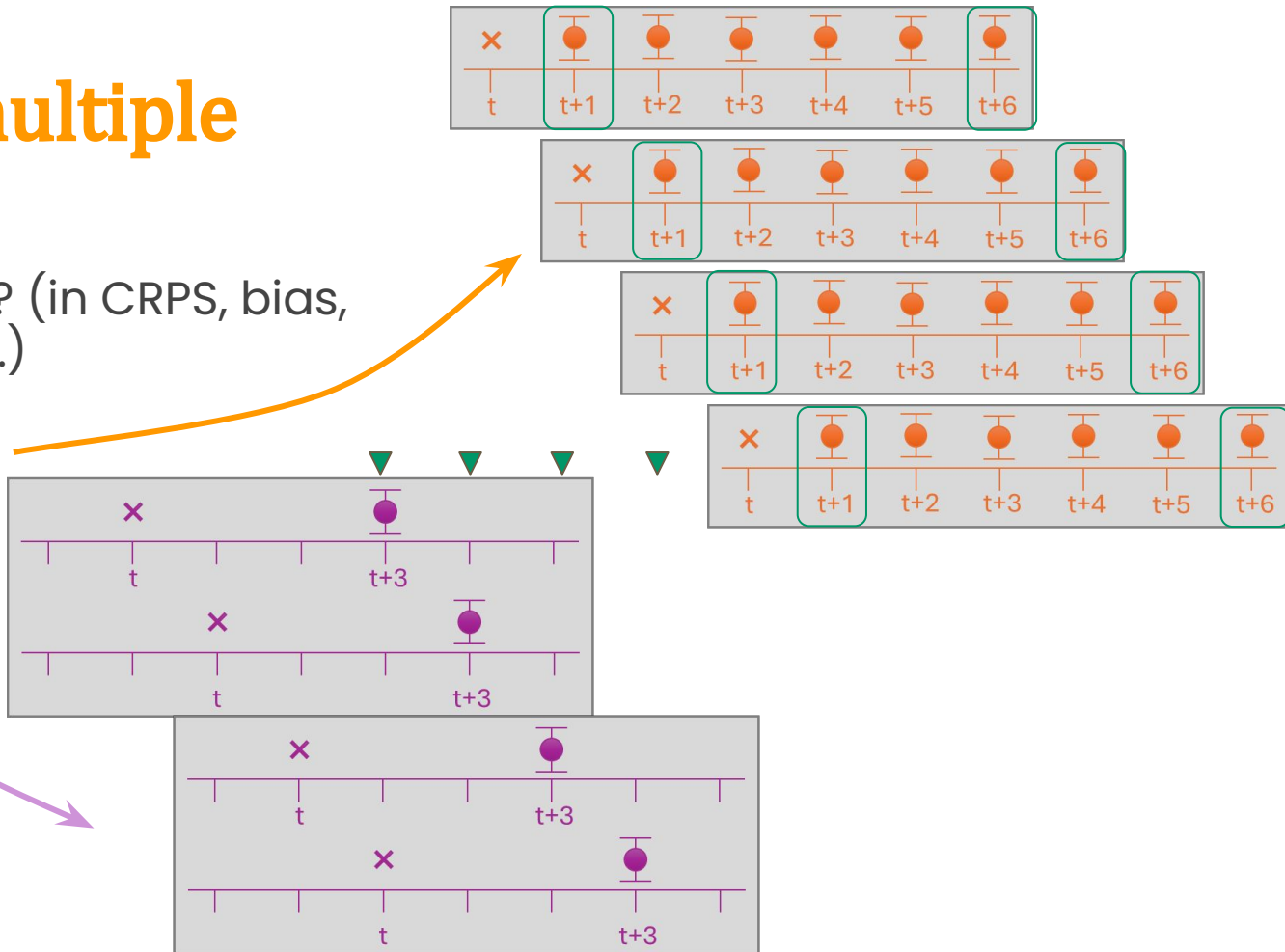
As with field ecology, a single sample does not a pattern make!



# Evaluating multiple forecasts

Do we see patterns? (in CRPS, bias, variance, reliability...)

- Across horizon
- Across time





# Evaluating multiple forecasts - skill scores

- Are we doing better than a baseline?
- Is this “complex” model performing better than a naive model?
- Useful in forecast model development
- Common baselines include climatology (day-of-year) and persistence (same as yesterday)
- Use the same metrics (RMSE, MAE, CRPS etc.)

$$1 - \frac{\text{Metric}_{\text{null}}}{\text{Metric}_{\text{forecast}}}$$

$$\frac{\text{Metric}_{\text{forecast}} - \text{Metric}_{\text{null}}}{\text{Metric}_{\text{opt}} - \text{Metric}_{\text{null}}}$$

$$\text{Metric}_{\text{forecast}} - \text{Metric}_{\text{null}}$$

# Questions and comments?

Part 2 materials:

<https://github.com/OlssonF/Forecast-evaluation-EFI25>

If you want to code along:

1. Download/fork/clone code
2. Install the packages

---

# References and further reading

Simonis, J. L., White, E. P., & Ernest, S. K. M. (2021). Evaluating probabilistic ecological forecasts. *Ecology*, 102(8), 1–8.

<https://doi.org/10.1002/ecy.3431>

Jolliffe, I. T., & Stephenson, D. B. (2012). *Forecast verification: a practitioner's guide in atmospheric science*. (Ian T. Jolliffe & D. B. Stephenson, Eds.) (2nd ed.). Oxford: Wiley Blackwell.

Gneiting, T., & Raftery, A. E. (2007). Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association*, 102(477), 359–378. <https://doi.org/10.1198/016214506000001437>

Smith, L. A., Suckling, E. B., Thompson, E. L., Maynard, T., & Du, H. (2015). Towards improving the framework for probabilistic forecast evaluation. *Climatic Change*, 132(1), 31–45. <https://doi.org/10.1007/s10584-015-1430-2>

Pappenberger, F., Ramos, M. H., Cloke, H. L., Wetterhall, F., Alfieri, L., Bogner, K., et al. (2015). How do I know if my forecasts are better? Using benchmarks in hydrological ensemble prediction. *Journal of Hydrology*, 522, 697–713.

<https://doi.org/10.1016/j.jhydrol.2015.01.024>

Wesselkamp, M., Albrecht, J., Pinnington, E., Castillo, W. J., Pappenberger, F., & Dormann, C. F. (2024). The ecological forecast horizon revisited: Potential, actual and relative system predictability. Retrieved from

<http://arxiv.org/abs/2412.00753>

Jacobs, B., Tobi, H., & Hengeveld, G. M. (2024). Linking error measures to model questions. *Ecological Modelling*, 487(July 2023), 110562. <https://doi.org/10.1016/j.ecolmodel.2023.110562>

Parker, W. S. (2020). Model Evaluation: An Adequacy-for-Purpose View. *Philosophy of Science*, 87(3), 457–477.

<https://doi.org/10.1086/708691>

# Applications with the NEON Challenge

Thomas, R. Q., Boettiger, C., Carey, C. C., Dietze, M. C., Johnson, L. R., Kenney, M. A., et al. (2023). The NEON Ecological Forecasting Challenge. *Frontiers in Ecology and the Environment*, 21(3), 112–113. <https://doi.org/10.1002/fee.2616>

Boettiger C, Thomas Q (2024). score4cast: Scoring Utilities for Forecast Competitions. R package version 0.0.0.9000. <https://github.com/eco4cast/score4cast>

Olsson, F., Carey, C. C., Boettiger, C., Harrison, G., Ladwig, R., Lapeyrolerie, M. F., et al. (2025). What can we learn from 100,000 freshwater forecasts? A synthesis from the NEON Ecological Forecasting Challenge. *Ecological Applications*, 35(1), 1–22. <https://doi.org/10.1002/eap.70004>

Wheeler, K. I., Dietze, M. C., LeBauer, D., Peters, J. A., Richardson, A. D., Ross, A. A., et al. (2024). Predicting spring phenology in deciduous broadleaf forests: NEON phenology forecasting community challenge. *Agricultural and Forest Meteorology*, 345, 109810. <https://doi.org/10.1016/j.agrformet.2023.109810>

# Ecological forecast evaluation in practice

## CRPS, log and other probabilistic/distributional forecast scores

Ouellet-Proulx et al., 2017. Water temperature ensemble forecasts: Implementation using the CEQUEAU model on two contrasted river systems. *Water (Switzerland)*, 9(7). <https://doi.org/10.3390/w9070457>

Olsson, F. et al., 2024. A Multi-Model Ensemble of Baseline and Process-Based Models Improves the Predictive Skill of Near-Term Lake Forecasts. *Water Resources Research*, 60(3). <https://doi.org/10.1029/2023WR035901>

Johansson, M. A., et al., 2019. An open challenge to advance probabilistic forecasting for dengue epidemics. *Proceedings of the National Academy of Sciences of the United States of America*, 116(48), 24268–24274. <https://doi.org/10.1073/pnas.1909865116>

Chan, K.H., et al., 2025. Data-driven approach to weekly forecast of the western flower thrips (*Frankliniella occidentalis* Pergande) population in a pepper greenhouse with an ensemble model. *Pest Management Science*. <https://doi.org/10.1002/ps.8713>

Dumandan, P.K.T., et al., 2024. Transferability of ecological forecasting models to novel biotic conditions in a long-term experimental study. *Ecology*, 105(11), p.e4406. <https://doi.org/10.1002/ecy.4406>

# Ecological forecast evaluation in practice

## Forecast predictive interval reliability and calibration

Gneiting, T. et al., 2007. Probabilistic Forecasts, Calibration and Sharpness. Journal of the Royal Statistical Society Series B: Statistical Methodology, 69(2), 243–268. <https://doi.org/10.1111/j.1467-9868.2007.00587.x>

Zwart, J. A. et al., 2023). Near-term forecasts of stream temperature using deep learning and data assimilation in support of management decisions. JAWRA Journal of the American Water Resources Association, 59(2), 317–337. <https://doi.org/10.1111/1752-1688.13093>

Thomas, R. Q. et al., 2020. A Near-Term Iterative Forecasting System Successfully Predicts Reservoir Hydrodynamics and Partitions Uncertainty in Real Time. Water Resources Research, 56, e2019WR026138. <https://doi.org/10.1029/2019WR026138>