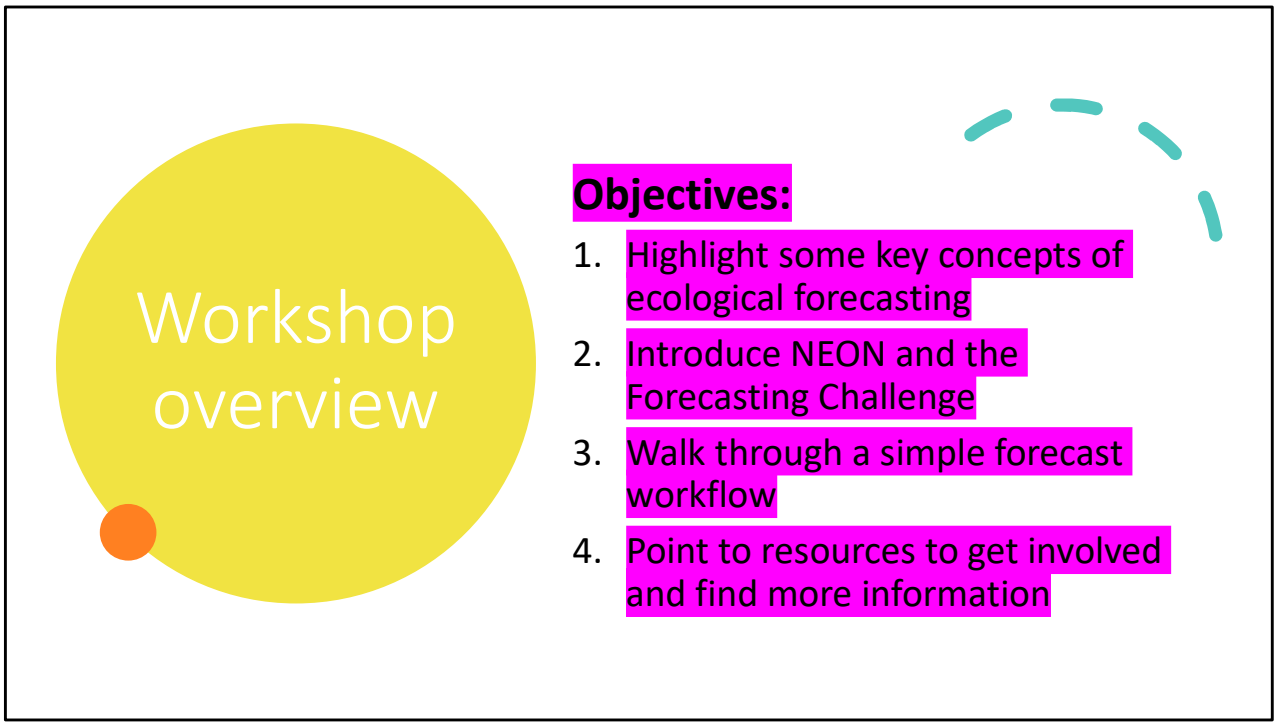
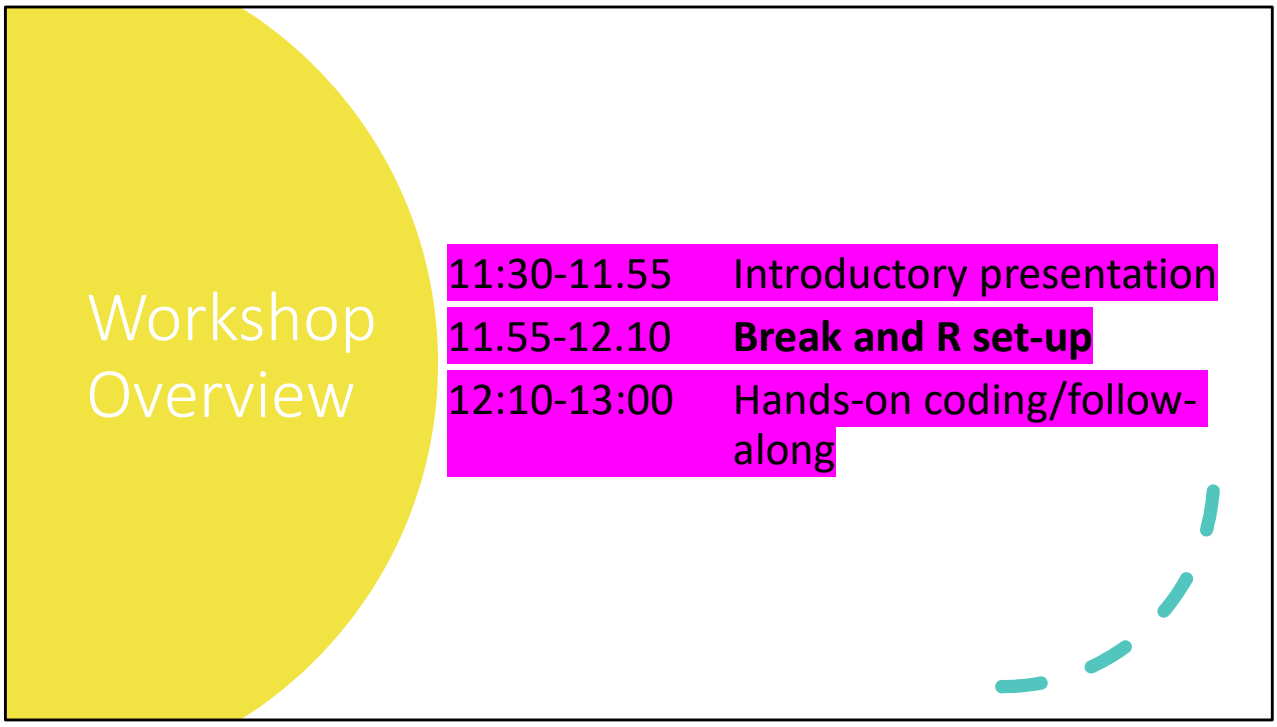
Can you predict the future? Introduction to the NEON Forecasting Challenge

Workshop lead:
Contact:

Text highlighted in pink should be modified where appropriate.
Slides relating specifically to the terrestrial theme are hidden but can be used in place of those about the aquatics theme.

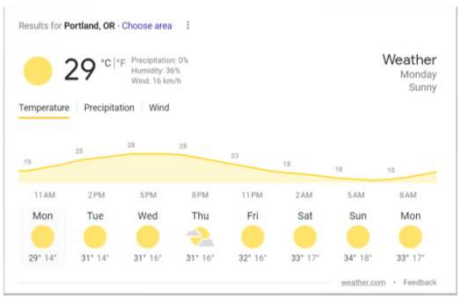


List your primary objective for the workshop. This might be introducing forecasting as a concept, developing forecast workflows, using forecast models, and/or automating forecasts



A rough schedule for the workshop. The Submit_forecast tutorial are designed to be completed in a 90-minute workshop

Why forecast?



How many people looked at the weather this morning? Or last weekend when you were packing for a weekend trip? Weather forecasts are also crucial for emergency planning with natural disasters such as hurricanes, forecasts for the wildfires and associated air quality. Similarly in ecology, these near term forecasts may prove to be equally valuable for managing ecosystems and informing public.

Figures show a Google 7 day weather forecast; a NOAA hurricane forecast for Hurricane Laura 2020; NICC 7-day fire potential outlook

Near-term, iterative, ecological forecasts

- **Near-term** = sub-daily to decadal timescales
- **Iterative** = process of repeatedly validating forecasts, updating model initial conditions and parameters, and issuing new forecasts as new data become available
- **Ecological forecast** = future predictions of physical, chemical, or biological variables *with quantified uncertainty*

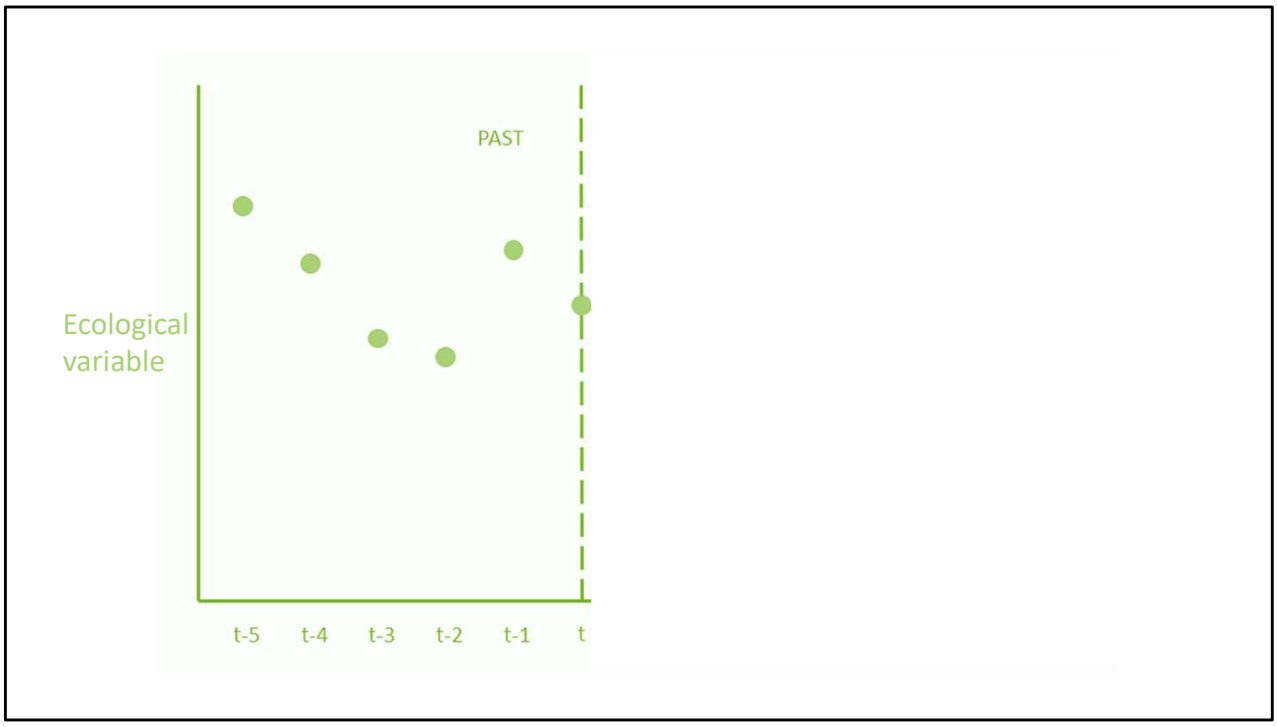
What is meant by a near-term, iterative, ecological forecast? It is useful to define what is and isn't meant when we use these terms (at least for these materials specifically.

Near-term – we are focusing on predictions occurring on a sub-daily to decadal time scale. These are not climate projections.
Iterative – the act of continually updating and rerunning a forecast workflow. We don't just make a single one-off forecast but are constantly evaluating, updating and generating new forecasts as more data are collected.
Ecological forecast – in this scenario we are making real *future* predictions of the environment. Importantly the forecast needs to include estimates of uncertainty. Because they are real-time forecasts they inherently have uncertainty that should be quantified and communicated as part of the prediction.

Examples:

1. Forecast of river dissolved oxygen concentration for the next 1-48 hours for fish stocking
2. 1-3 month ahead predictions of the % chance of leaf fall to estimate peak leaf-peeping
3. Forecasts of tick abundance for the next 1-30 days in a popular hiking area

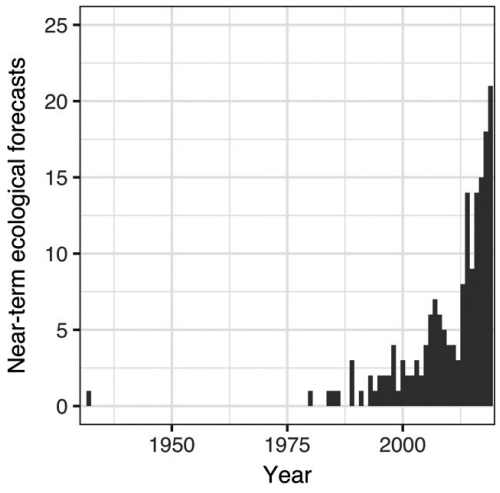
Some examples of some near term, iterative, ecological forecasts.
Note these relate to the NEON Forecasting Challenge themes (aquatics, ticks, phenology)



Example of what we mean by a forecast and how you could visualize it. Past observation as points, future prediction (generated at dotted line), with uncertainty

Ecological forecasting

An emerging and growing field

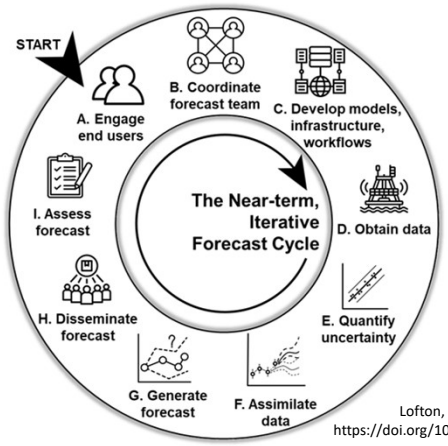


Lewis et al., 2021
(https://doi.org/10.1002/eap.2500)

Ecological forecasting is a relatively new field but one that has the potential to have both applied and fundamental outcomes.
Lots to learn about predictability, best methods etc.

Forecasting Challenges

Forecasting is challenging!



Lofton, M. E *et al.* (2023)
https://doi.org/10.1111/gcb.16590

A Challenge to catalyze progress

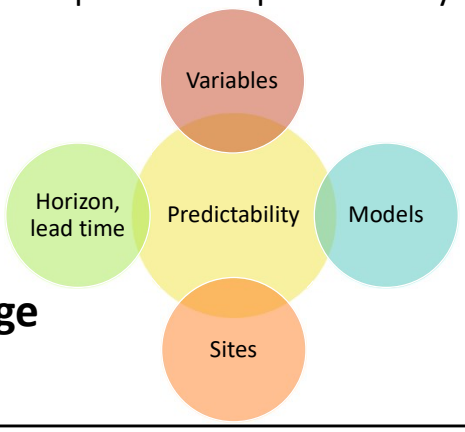
The endeavour of producing a real-time forecast is not easy given the infrastructure needed to collect, disseminate data, develop and deploy models, evaluate and distribute forecasts to end-users. In addition the iterative nature of the forecast cycling requires these steps occur in a reproducible and repeatable way.

In other fields such as economics and computing competitions and challenge have been used as a way to forward the field and provide cohesion amongst researchers. An organising venture/enterprise for the community

Why a forecasting challenge? – the power of many forecasts!

1. A community of forecasting
   - Standards
   - Development of tools and infrastructure
   - A forecasting **platform**

2. Answer questions of predictability



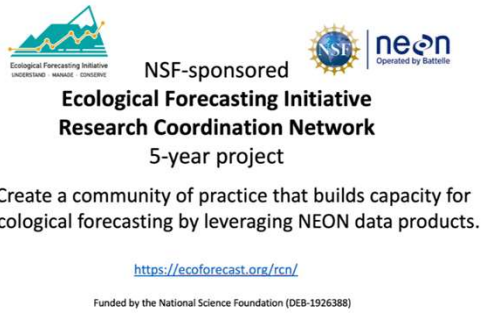
The **EFI-NEON Forecasting Challenge** was born (2021)!

A Challenge has two potential values:

1. Generates a community of forecasters with common standards of production, submission, evaluation etc.
2. Start to answer questions of predictability across a range of scales (models, sites, variables). NEON data is perfect for this

Ecological Forecasting Initiative Research Co-ordination Network

- EFI RCN Goals
  - lower barriers
  - community building
  - infrastructure
  - platform development

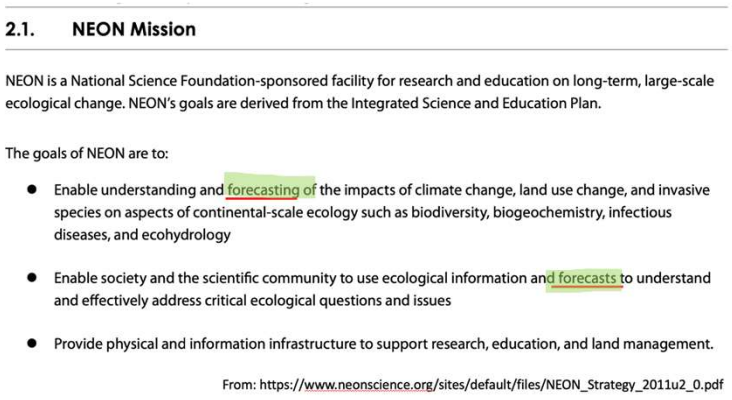


EFI RCN supported by NSF grant DEB-1926388

Challenge coordinated by the Ecological Forecasting Initiative Research Coordination Network as a means to engage more people
Partnered with NEON for the Challenge

What is neon?

- The National Ecological Observatory Network (NEON) is a **continental-scale observation facility**
- To collect **long-term open access ecological data**
- **47 terrestrial and 34 aquatic sites**



2.1. NEON Mission

NEON is a National Science Foundation-sponsored facility for research and education on long-term, large-scale ecological change. NEON's goals are derived from the Integrated Science and Education Plan.

The goals of NEON are to:

- Enable understanding and forecasting of the impacts of climate change, land use change, and invasive species on aspects of continental-scale ecology such as biodiversity, biogeochemistry, infectious diseases, and ecohydrology
- Enable society and the scientific community to use ecological information and forecasts to understand and effectively address critical ecological questions and issues
- Provide physical and information infrastructure to support research, education, and land management.

From: https://www.neonscience.org/sites/default/files/NEON_Strategy_2011u2_0.pdf

The Challenge leverages the continental scale ecological observatory network (NEON) that collects data across the US from 81 different sites.

A screenshot from the NEON justification in early 2000s shows that "forecast ecosystems" is key goal within the NEON mission.

What is the EFI-NEON Challenge?

"A platform for the community to make predictions of conditions at NEON sites before the data are collected"

- All 81 sites
- 6 themes

The Challenge covers 5 themes across ecology: Ticks, beetles, phenology, terrestrial daily, terrestrial subdaily, aquatics
Created a platform for the ecological forecasting community to make

What is the EFI-NEON Challenge?

"A platform for the community to make predictions of conditions at NEON sites before the data are collected"

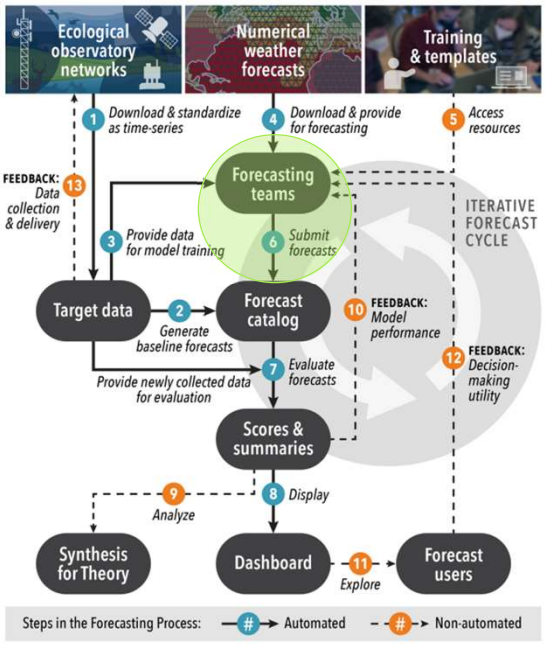
- All 81 sites
- 6 themes
- > 15,000 forecast submitted!

Thomas, 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

The Challenge has created a platform for the ecological forecasting community to make forecasts and do synthesis
More information on the Challenge at Thomas et al 2023.

Updating the numbers can be found here: https://projects.ecoforecast.org/neon4cast-ci/ (look for

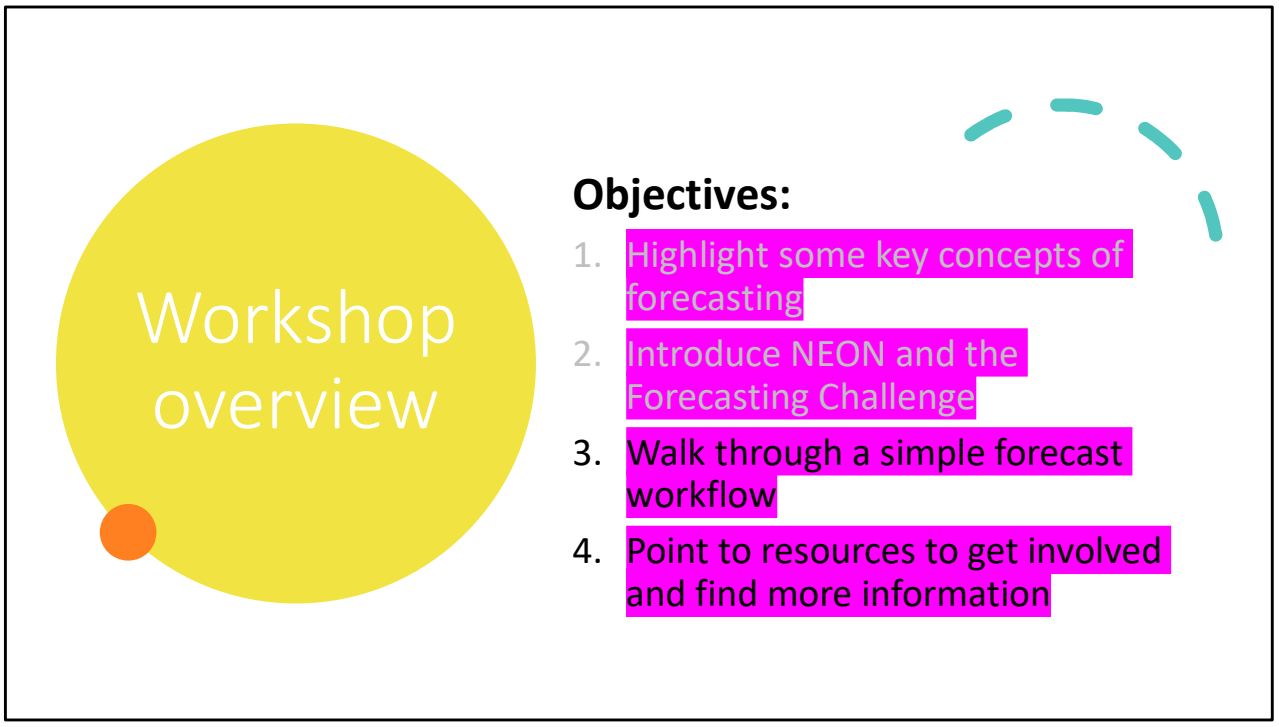
Challenge overview



Thomas, 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

Use this slide to walk through the steps of the challenge. The materials here focus on Forecasting Teams submitting forecasts. The CI implemented by organisers has taken care of a lot of the rest of the workflow

Full paper at QR code – great for an overview (pre-reading etc.)!



Time for questions and discussion

Hands-on workshop:

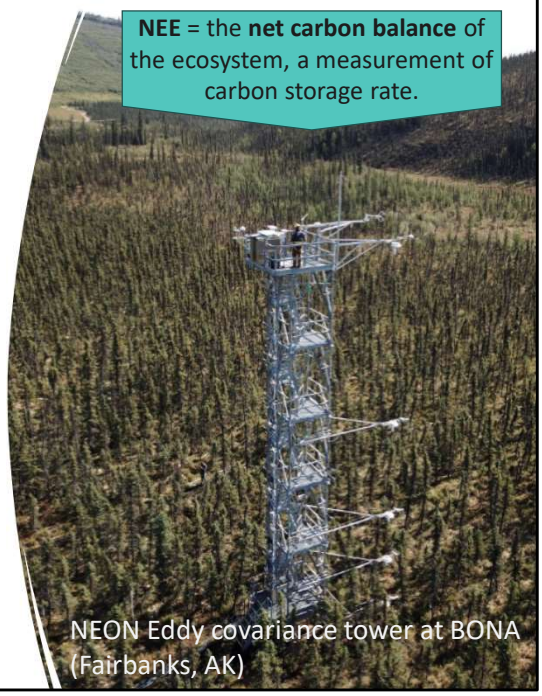
- **Aquatics theme – Can we predict how water temperature will change over the next month?**
  - Water temperature in lakes
  - NEONs water temperature data product (DP1.20264.001)
  - 30 day forecast horizon
  - Data latency of 2-3 days
- **Simple baseline** model to build on

An overview of the data and questions we are working with in this tutorial.
Why water temperature?
We are using a simple model (not necessarily a good forecast) to demonstrate a full workflow (from data to models, uncertainty, and submission)

Data that we will be using is collected by automated sensors at the sites (NEON Aquatic Instrument System (AIS), see https://www.neonscience.org/field-sites/about-field-sites). Data collected and quality checked by NEON are further processed to the targets format which is distributed to teams

Hands-on workshop:

- **Terrestrial daily theme – Can be predict how much C is going to be absorbed by ecosystems?**
  - net ecosystem exchange (NEE)
  - NEONs eddy covariance data product (DP4.00200.001)
  - 30 day forecast horizon
  - Data latency of 5 days
- **Simple baseline** model to build on

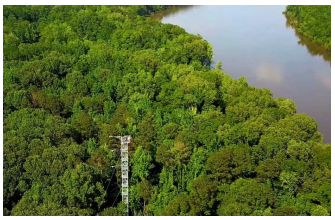


NEON Eddy covariance tower at BONA (Fairbanks, AK)

IF you are focusing on terrestrial Data collected across all 47 terrestrial
Why do we care about terrestrial theme? Carbon fluxes important – natural climate solutions
How much C is being absorbed by ecosystems?

We are using a simple model (not necessarily a good forecast) to demonstrate a full workflow (from data to models, uncertainty, and submission)

Diversity of NEON sites

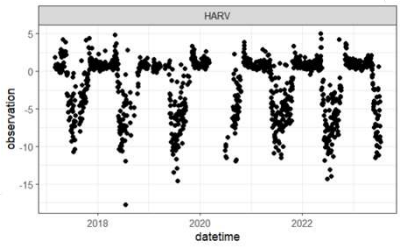






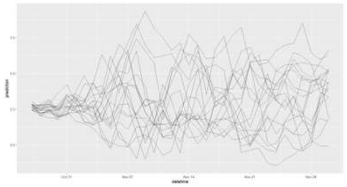
Learn more: https://www.youtube.com/watch?v=CR4Anc8Mkas

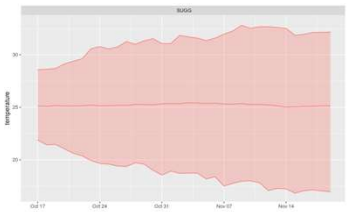
Some forecast terminology:



- Targets – water temperature

- Uncertainty – forecasts must include an estimate of uncertainty.
  The uncertainty can be represented using different model runs (**ensemble members**) or the statistics of the forecast (**mean and standard deviation**).



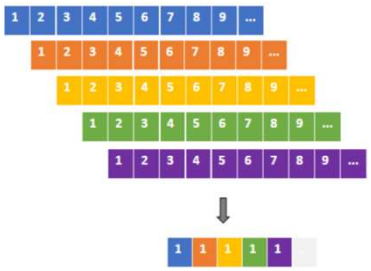


Some terminology that will come up in the workshop. Figure shows targets or observations (top). Not always raw data. Sometimes targets are derived from observations (subset, averaging etc.) But this is what we are trying to forecast and what will be used in evaluation.

Forecasts are inherently uncertain and there needs to be some estimate of what this is. This can be presented in one of two ways – an ensemble forecast (middle) that has multiple simulations of the future or by reporting the mean and standard deviation (bottom)

Some more forecast terminology:

- NOAA data – National Oceanic and Atmospheric Administration weather forecasts
- 3 NOAA forecast data products available in neon4cast:
  - Stage_1: **raw forecasts** from NOAA.
    30 member ensemble forecast
  - Stage_2: processed from stage_1 Recommended for **future forecasts. Hourly inputs**
  - Stage_3: the **historic data product**. A 'stacked' data set taking every 1 day ahead forecast.
    Useful for model training/calibration.



We will be using NOAA weather forecast products in our predictions. These forecasts have multiple simulations so teams can include this as a source of uncertainty in their ecological forecasts. Challenge organisers have compiled 3 data products that all teams can use to drive models and ensure consistency between ecological forecasts. These are the stage 1 and 2 data – the raw and processes ensemble weather forecasts. They have also developed an estimate of historical conditions called a stacked data product which can be used to train models. Takes all the 1 day ahead forecasts as a good estimate of observed conditions.

This gives a consistent data product that can be used for training and then forecasting (not switching between observations and models which could be biased, ensures consistent bias between training and forecasting).

A little more forecast terminology:

Scores – a means to assess forecast skill. The Challenge uses the Continuous Rank Probability Score (crps). Uses both the **accuracy** (mean) and the **precision** (sd) of the forecast.

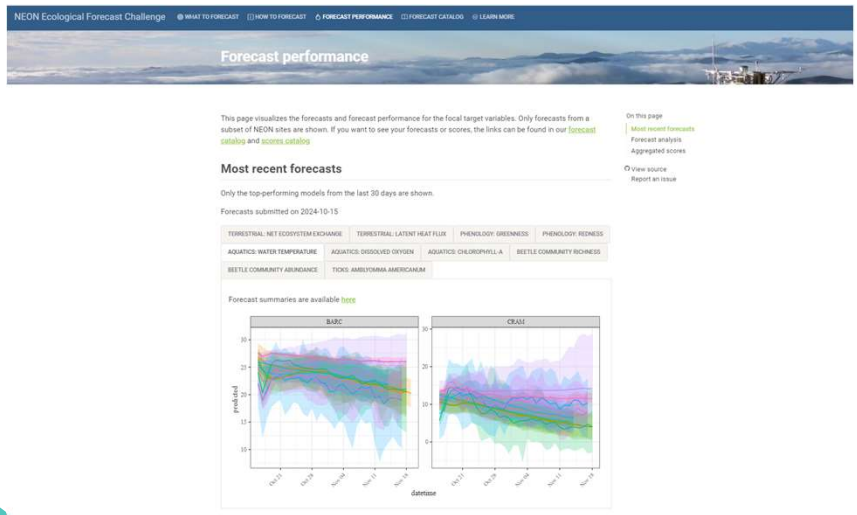
Scores → Dashboard → Users!

Read more: https://projects.ecoforecast.org/neon4cast-docs/Evaluation.html

Some other terms that I might use but you don't need to worry too much about are the scores – this is how the submitted forecasts are evaluated, how well did we do at forecasting the future. The method used by the Challenge incorporates the accuracy and the precision of the forecast – so how close where the predictions to the observation and how certain were we about that.

Scores are automatically generated and pushed to the dashboard for teams to see.

Read more about scores here https://projects.ecoforecast.org/neon4cast-docs/Evaluation.html



https://projects.ecoforecast.org/neon4cast-ci/

The Challenge includes a dashboard where all submitted forecasts are visualized. Once you have made a submission you can go and see how it performs relative to other submissions and baseline models.

A little more forecast terminology:

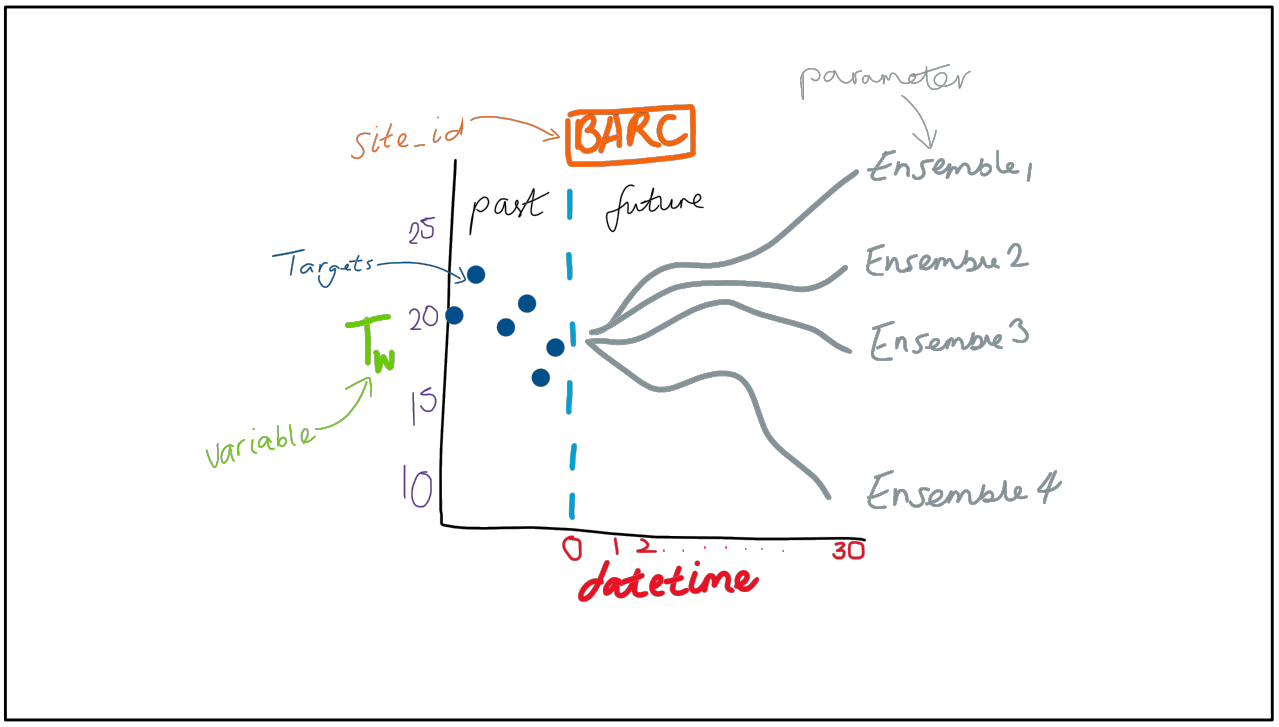
Standards - Help maintain consistency in forecast generation, submissions and scoring

Need to submit a forecast in a standardized format

- file format (csv or NetCDF)
- file name format ([theme]-[reference_datetime]-[team_name].csv)
- specific column names
- column format (datetime/character/integer/etc.)

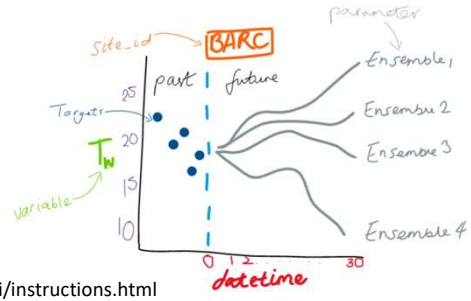
The standards are a set of requirements for formats etc. that help consistency and the automate methods – things need to have particular names and formats.

See here https://projects.ecoforecast.org/neon4cast-docs/Submission-Instructions.html



Before starting on the hands-on portion of the tutorial it's a good idea to visualize what a forecast that we generate could look like based on the Challenge's required standards. This drawing shows the targets are observations (blue points) of our chosen variable (water temperature, green). The forecast is for a particular NEON location (4-character NEON code, orange). The date that the forecast is generated is the vertical blue line, between past and future and is termed the reference datetime. If a forecast was generated today, then the reference datetime would be today's date and each of the associated predictions would be for a particular date (red). To represent uncertainty this forecast uses multiple predictions or ensemble members. The identify of the ensemble member is given by the parameter number – in this case 1 to 4.

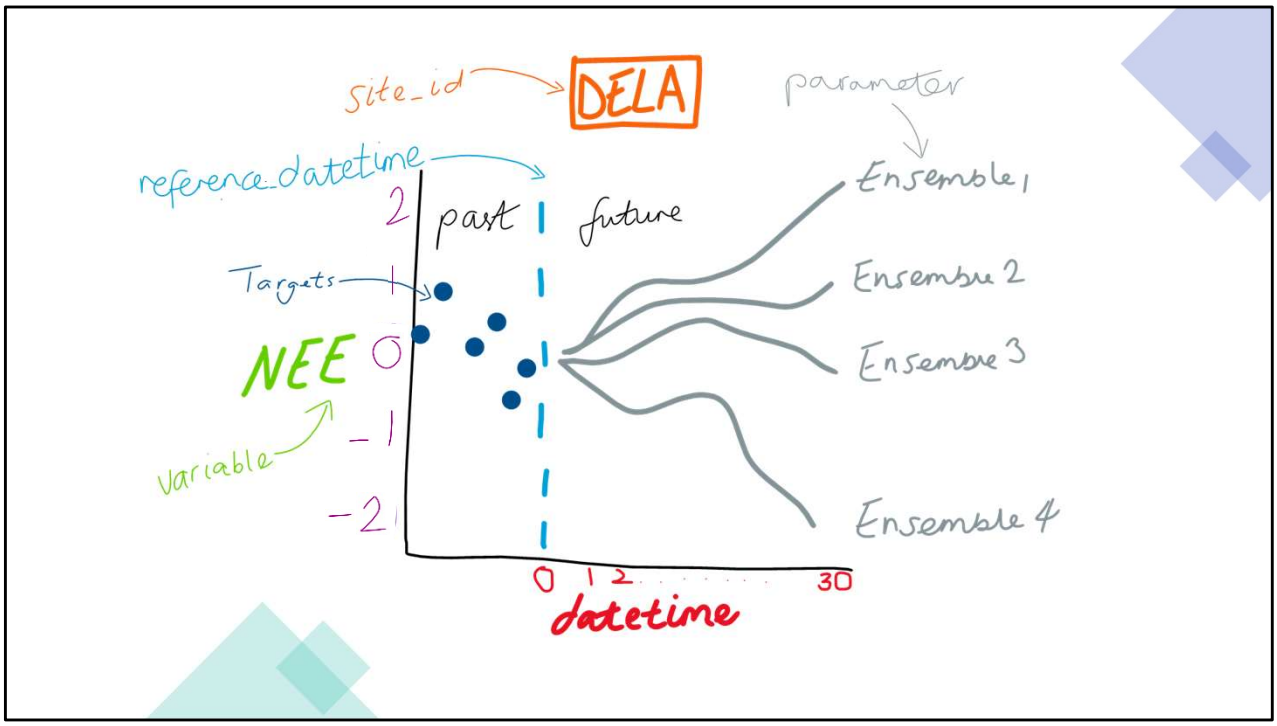
datetime	reference_datetime	site_id	family	parameter	variable	prediction	model_id
2023-01-12	2023-01-11	BARC	ensemble	1	temperature	22.63563	test_mod
2023-01-12	2023-01-11	BARC	ensemble	2	temperature	26.75148	test_mod
2023-01-12	2023-01-11	BARC	ensemble	3	temperature	24.65157	test_mod
2023-01-12	2023-01-11	BARC	ensemble	4	temperature	25.1389	test_mod
...	...	...	...	...	...	...	test_mod
2023-02-10	2023-01-11	BARC	ensemble	1	temperature	19.40379	test_mod
2023-02-10	2023-01-11	BARC	ensemble	2	temperature	24.89667	test_mod
2023-02-10	2023-01-11	BARC	ensemble	3	temperature	25.98961	test_mod
2023-02-10	2023-01-11	BARC	ensemble	4	temperature	26.40593	test_mod



https://projects.ecoforecast.org/neon4cast-ci/instructions.html

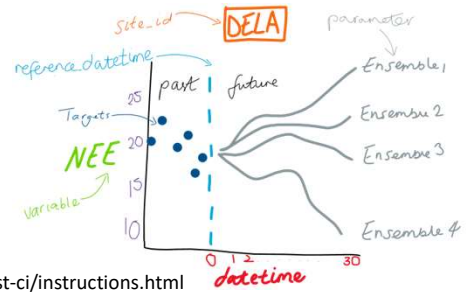
This drawing would be represented in a data frame that looks like this. This includes the require columns names and formats. The family and parameter columns are crucial for describing the uncertainty representation in the forecast. The family describes how the uncertainty is represented in this case an ensemble but if the uncertainty is drawn from a distribution (e.g. normal, binomial etc.) that distribution would be listed in the family. For an ensemble forecast parameter is the ensemble number, but for a distributional forecast these would be the parameters of that distribution (e.g. for a normal distribution it would be mu (mean) and sigma (sd)). A full description of the required columns and possible families can be found https://projects.ecoforecast.org/neon4cast-ci/instructions.html

Update the dates to more recent to reflect what participants might see.



Before starting on the hands-on portion of the tutorial it's a good idea to visualize what a forecast that we generate could look like based on the Challenge's required standards. This drawing shows the targets are observations (blue points) of our chosen variable (NEE, green). The forecast is for a particular NEON location (4-character NEON code, orange). The date that the forecast is generated is the vertical blue line, between past and future and is termed the reference datetime. If a forecast was generated today, then the reference datetime would be today's date and each of the associated predictions would be for a particular date (red). To represent uncertainty this forecast uses multiple predictions or ensemble members. The identify of the ensemble member is given by the parameter number – in this case 1 to 4.

datetime	reference_datetime	site_id	family	parameter	variable	prediction	model_id
2023-08-02	2023-08-01	DELA	ensemble	1	nee	5.551	test_mod
2023-08-02	2023-08-01	DELA	ensemble	2	nee	4.547	test_mod
2023-08-02	2023-08-01	DELA	ensemble	3	nee	2.227	test_mod
2023-08-02	2023-08-01	DELA	ensemble	4	nee	3.214	test_mod
…	…	…	…	…	…	…	test_mod
2023-09-02	2023-08-01	DELA	ensemble	1	nee	6.852	test_mod
2023-09-02	2023-08-01	DELA	ensemble	2	nee	2.247	test_mod
2023-09-02	2023-08-01	DELA	ensemble	3	nee	5.961	test_mod
2023-09-02	2023-08-01	DELA	ensemble	4	nee	4.593	test_mod



https://projects.ecoforecast.org/neon4cast-ci/instructions.html

In a csv format that meets standards it might look like this. One important thing to understand is the idea of a reference_Datetime or the date when the forecast was generated. Here you can see the forecast was generated on 1st August and goes out 30 days into the future. Family and parameter describe the type of forecast it is – family is ensemble and the parameter is the ID of each ensemble member

This drawing would be represented in a data frame that looks like this. This includes the require columns names and formats. The family and parameter columns are crucial for describing the uncertainty representation in the forecast. The family describes how the uncertainty is represented in this case an ensemble but if the uncertainty is drawn from a distribution (e.g. normal, binomial etc.) that distribution would be listed in the family. For an ensemble forecast parameter is the ensemble number, but for a distributional forecast these would be the parameters of that distribution (e.g. for a normal distribution it would be mu (mean) and sigma (sd)). A full description of the required columns and possible families can be found https://projects.ecoforecast.org/neon4cast-ci/instructions.html

Basic workflow to submit a forecast

1. Read EFI-NEON Challenge documentation (neon4cast.org)
2. Investigate the forecast target variables
3. Build/apply your model!
4. Produce forecast of future conditions – SUBMIT TO THE CHALLENGE!
5. Register, complete model description, and submit your forecasts
6. Wait for the scores to come in and revel in the glory of predicting the future (~5-day before first evaluation)
7. Use new data to update the model
8. Submit another forecast! And another…!

Workflow automation

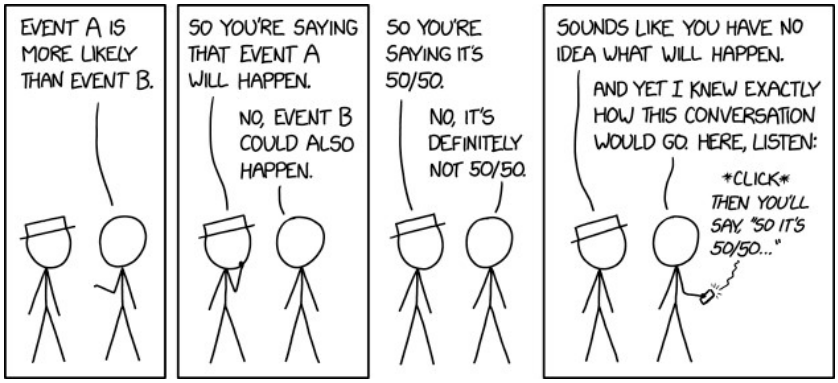
See detailed instructions here https://projects.ecoforecast.org/neon4cast-ci/instructions.html

To put this together this is what a simple forecasting workflow might look like to submit to the NEON Challenge. Lots of great documentation and templates and tools to help you get started

Get involved in the iterative nature of forecasting, submit/improve your forecast every day. Use data assimilation techniques to modify starting conditions

https://projects.ecoforecast.org/neon4cast-ci/instructions.html

Questions?

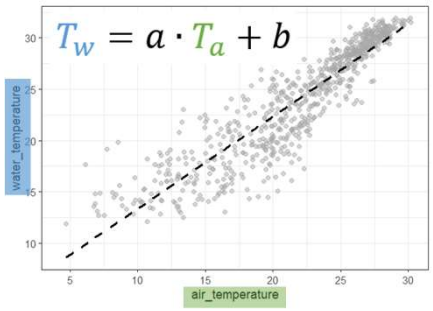


https://xkcd.com/2370

Let's forecast!

1. Follow-along R markdown
   - Forecasting water temperature using a Linear model with air temperature

2. Modify the model and submit your forecast!
   - More/other covariates
   - Different model structures
   - Other variables



A preview of the forecasting model to be used – forecasting water temperature as a function of air temperature using weather forecasts

Let's forecast!

1. Follow-along R markdown
   - Forecasting Net Ecosystem Exchange using a Linear model with air temperature and short-wave radiation

$$NEE = c + b \cdot AT + a \cdot SWR$$

Or just watch along!

2. Modify the model and submit your forecast!
   - More/other covariates
   - Different model structures
   - Other variables

A preview of the forecasting model to be used – forecasting NEE as a function of air temperature and short wave radition using weather forecasts

Thank you for attending!

Big thanks to the EFI-NEON team - especially **Quinn Thomas** (Virginia Tech) and **Carl Boettiger** (UC-Berkley) developers of the cyberinfrastructure underpinning the Challenge.

For questions contact
eco4cast.initiative@gmail.com

Visit ecoforecast.org & neon4cast.org