Climate and Oceans Support
Program in the Pacific

ACCESS-S Workshop

MODULE: ACCESS-S Model Skill

Topics in this module

- Why verify?
- How to verify seasonal climate forecasts
- Linear Error in Probability Space (LEPS)
- ERA5 reanalysis for verification
- Climate driver skill
- SPCZ location in ACCESS
- Tropical cyclone skill

Expected learning outcomes:

- Understanding of how a seasonal climate forecast is verified
- Understanding the methods used to verify ACCESS-S forecasts
- Understanding the datasets used to verify ACCESS-S forecasts

Why verify?

A forecast is not complete until you find out whether the forecast was successful.

Three important reasons to verify forecasts:

1. to ***monitor*** forecast quality - how accurate are the forecasts and are they improving over time?
2. to ***improve*** forecast quality - the first step toward getting better is discovering what you're doing wrong.
3. to ***compare*** the quality of different forecast systems - to what extent does one forecast system give better forecasts than another, and in why is that system better?

Verification

- How do we decide if a forecast is ***correct***?
- How do we decide if a set of forecasts are ***correct*** often enough to be ***good***?
- How do we decide if climate forecasts are ***valuable***?

Verification ~~can be~~ is complicated

When is a forecast correct?

Climate and Oceans Support Program in the Pacific

Tomorrow it will rain....

Observed: Rain	<u>correct or incorrect?</u>	**Correct**
Observed: No Rain	<u>correct or incorrect?</u>	**Incorrect**

Tomorrow there is a 70% chance of rain....

Observed: Rain	<u>correct or incorrect?</u>	**Correct**
Observed: No Rain	<u>correct or incorrect?</u>	**Also Correct**

Measuring Accuracy (reliability)

Unless the probability is 0 or 100%, a single probability forecast is *always* correct.

e.g., 90% chance of rain implies 10% chance of no rain.

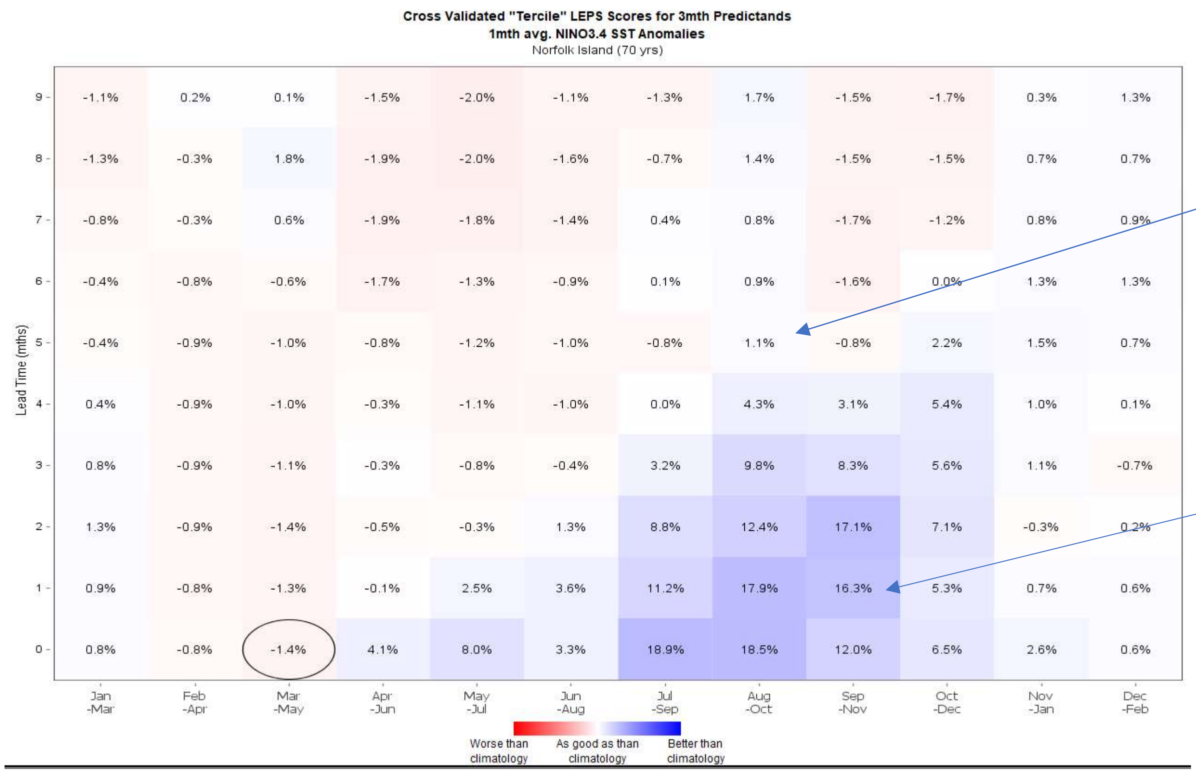
However, the forecast may have the wrong confidence – this is reliability.

When a forecaster says there is a high probability of rain tomorrow, it should rain more frequently than when the forecaster says there is a low probability of rain.

What is Skill? Not the same as accuracy

Climate and Oceans Support Program in the Pacific

Skill measures how much better a forecast is than some benchmark – e.g. a climate forecast of 33:33:33%
Often use Linear Error in Probability Space (LEPS).



Low skill – not much different from a benchmark

High skill – much better than a benchmark

Climate and Oceans Support Program in the Pacific

Why might a forecast have *Low Skill*?

- If the model cannot capture mean climate features, such as the MJO, ITCZ or SPCZ
- If the model cannot capture simulation of the climate variability such as the advance and retreat of monsoons
- Predictive skill tends to decrease with lead times
- Where there is **low skill**, forecasts are **not usually emphatic**
- Where there is low skill, **forecasts tend to give climatology** (e.g. 50:50% above/below median or 33:33:33 terciles

Metrics of verification

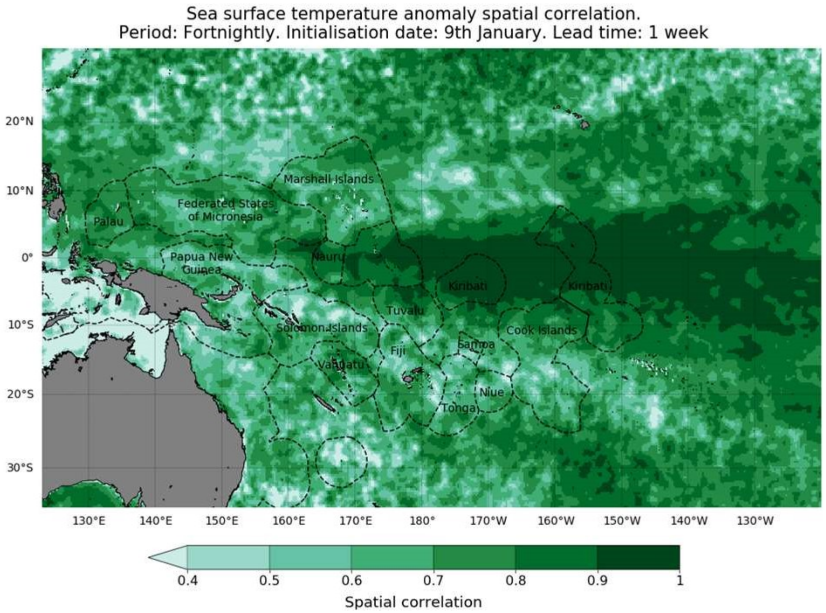
- Accuracy
- Probability of detection (hit rate)
- False alarm ratio
- Probability of false detection (false alarm rate)
- Threat score (critical success index)
- Hanssen and Kuipers discriminant (true skill statistic, Peirce's skill score)
- Heidke skill score
- Mean absolute error
- **Root mean square error**
- **Linear error in probability space (LEPS)**
- **Anomaly correlation**
- **Brier score**
- Reliability diagram
- Relative Operating Characteristic (ROC)
- Ranked probability skill score
- … …

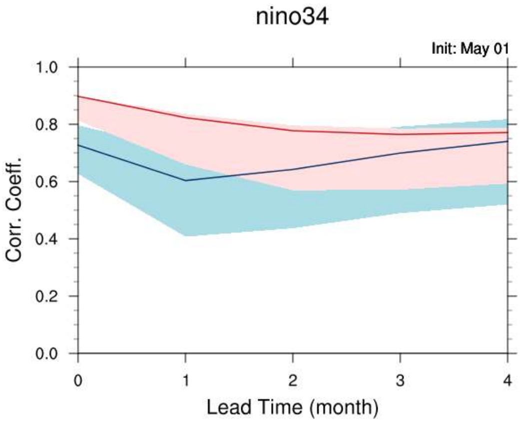
  – More details at http://www.cawcr.gov.au/projects/verification/

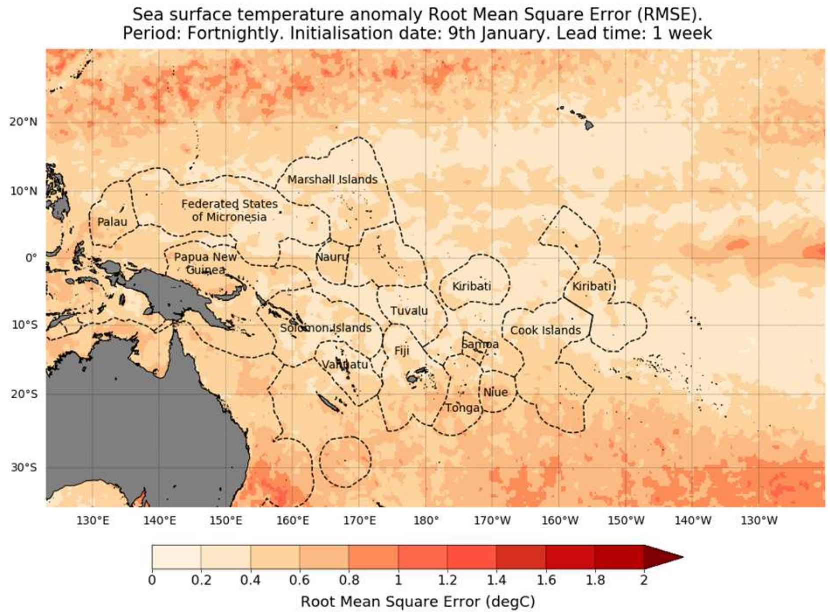
Hindcast skill: How good is ACCESS_S at prediction?

Climate and Oceans Support Program in the Pacific

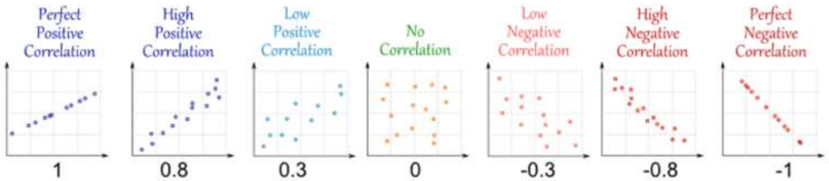
Different types of forecasts need different methods of verification







Correlation coefficient: measures how strong a relationship is between the model and real world over time



Root Mean Squared Error: often used alongside correlation, it's a measure of the average magnitude of the forecast error, *i.e.* how concentrated the data is around the line of best fit between the model and observed.

Hindcast skill: How good is ACCESS-S at prediction?

Climate and Oceans Support Program in the Pacific

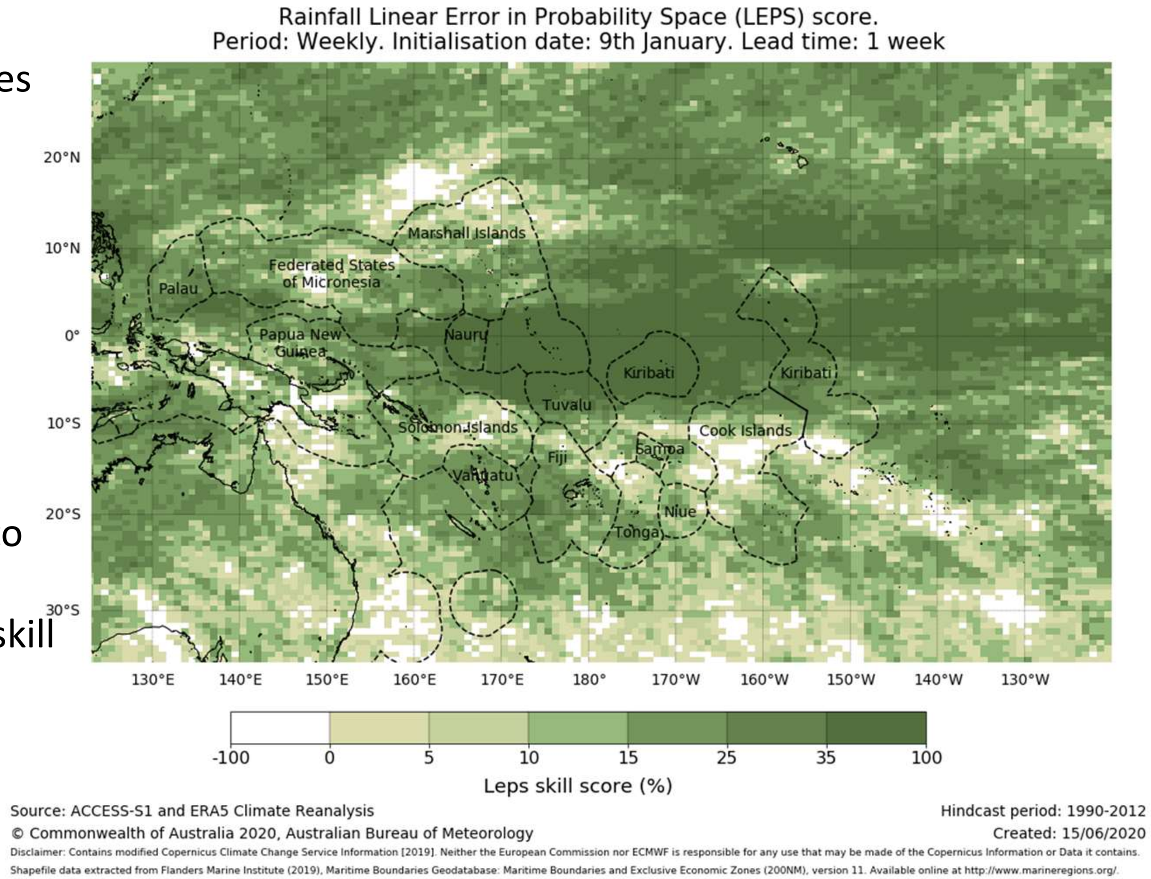
Different types of forecasts need different methods of verification

LEPS: measures the error in probability space, so it does not discourage forecasting extreme probabilities if they are warranted. The most useful for many aspects of probability forecasting, but it's more complicated than other metrics.

Darker green regions show the model predicted the correct tercile more often when compared to ERA5 'observations' for that same period

This indicates regions where Real-time skill is likely to be highest, however real-time skill should be monitored separately as it may not match hindcast skill exactly

https://www.researchgate.net/publication/253615254_Revised_LEPS'_Scores_for_Assessing_Climate_Model_Simulations_and_Long-Range_Forecasts

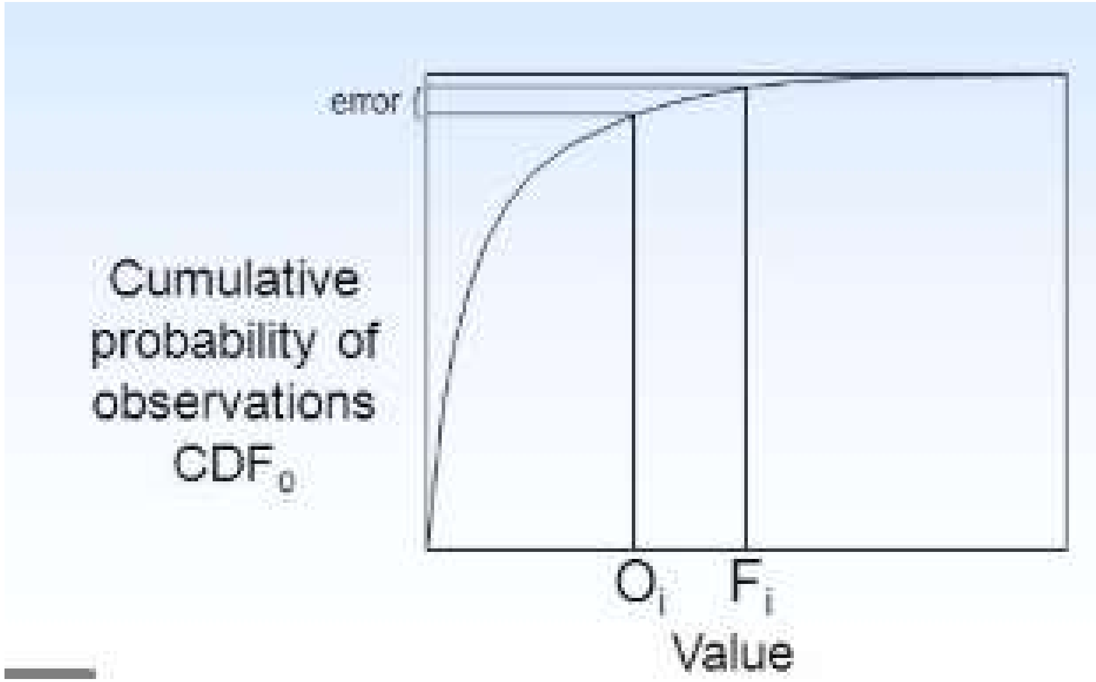


Linear Error in Probability Space (LEPS)

Advantages:
- Rewards emphatic forecasts
- Valid across all categories
- Can be mapped
- Rewards difficult/extreme forecasts

Disadvantages:
- Difficult to understand/not intuitive
- Difficult to calculate

- Measures Probability error – does not penalise more extreme forecasts when they are justified
- Errors in the tail of the distribution are penalized less than errors in the centre of the distribution



Another way of looking at it…Cumulative Probability Distribution

Climate and Oceans Support Program in the Pacific

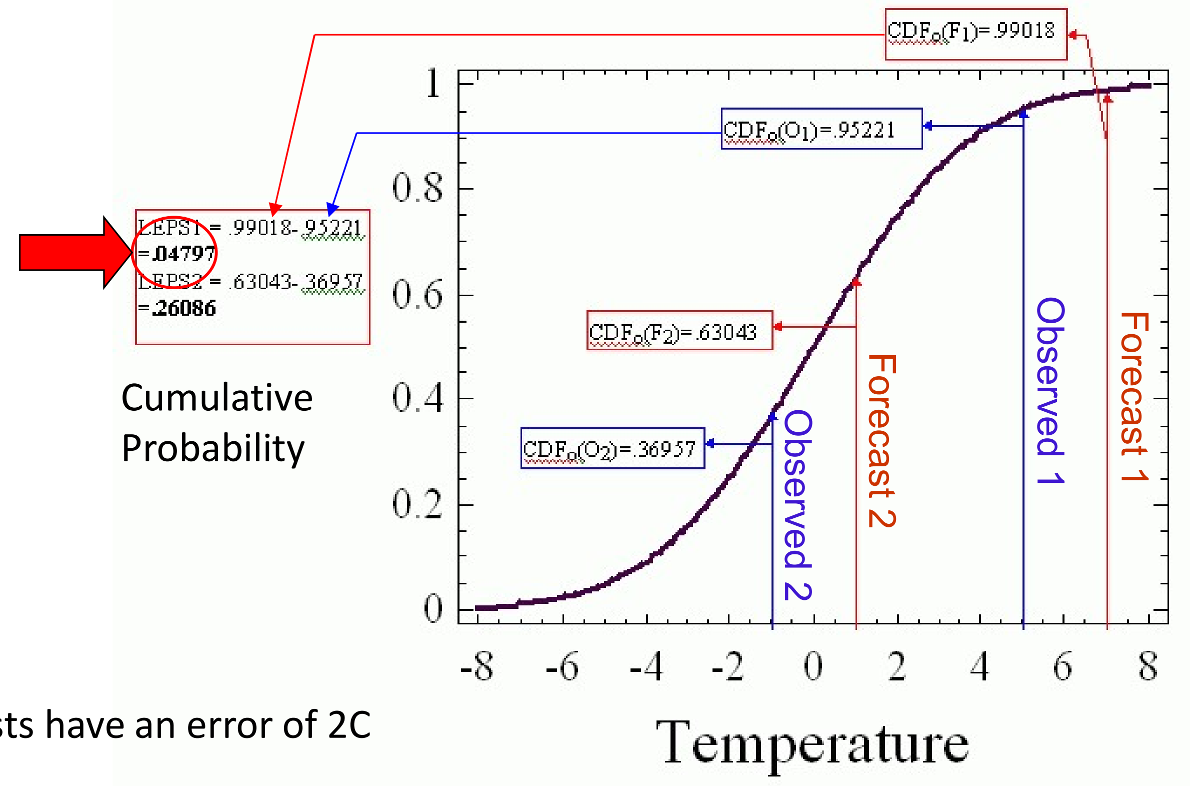
Remember- LEPS stands for "*Linear* ***Error*** *in* ***Probability*** *Space*"!

Penalize errors in terms of the distance between forecasts and observations in cumulative distribution space

It gives relatively more penalty when forecasting events around average values, but relatively higher scores and less penalty for forecasts of extreme events

Smaller **Error** for Forecast 1.

Note: This is not a LEPS "Score"



Note, both forecasts have an error of 2C

LEPS Skill assessment – categorical forecasts

Climate and Oceans Support Program in the Pacific

Observation	Prediction: Tercile 1	Prediction: Tercile 2	Prediction: Tercile 3
Tercile 1	8/27	–1/27	–7/27
Tercile 2	–1/27	2/27	–1/27
Tercile 3	–7/27	–1/27	8/27

LEPS Scoring Table for Terciles

Example: A prediction of Tercile 2, with a verifying observation of Tercile 3, will score –1/27. This is for a **categorical forecast**. Similarly, a prediction and observation of Tercile 3 will score 8/27.

Adding our scores over many forecasts provides information about the skill of our forecast system.

Very many (i.e. hundreds) of **random predictions will add to zero.** Why? Because the scores in each row or column of the table add to zero.

Zero is therefore our **baseline** or **climatological** LEPS score.

LEPS Skill assessment – Probability Forecasts

Climate and Oceans Support Program in the Pacific

Observation	Prediction: Tercile 1	Prediction: Tercile 2	Prediction: Tercile 3
Tercile 1	8/27	–1/27	–7/27
Tercile 2	–1/27	2/27	–1/27
Tercile 3	–7/27	–1/27	8/27

LEPS Scoring Table for Terciles

Suppose the seasonal forecast **probabilities** were **P1(34.2%)**, **P2(31.7%)** and **P3(34.1%)** and the observed category was tercile 3. The LEPS score is:
(**P3**x8/27) – (**P2**x1/27) – (**P1**x7/27) = **S**
(0.341*8/27**)-(0.317***1/27**)-(0.342***7/27**)=0.00063**

A perfect forecast (P3 = 1.0) will score 8/27, in this instance. Our perfect score = **PF**

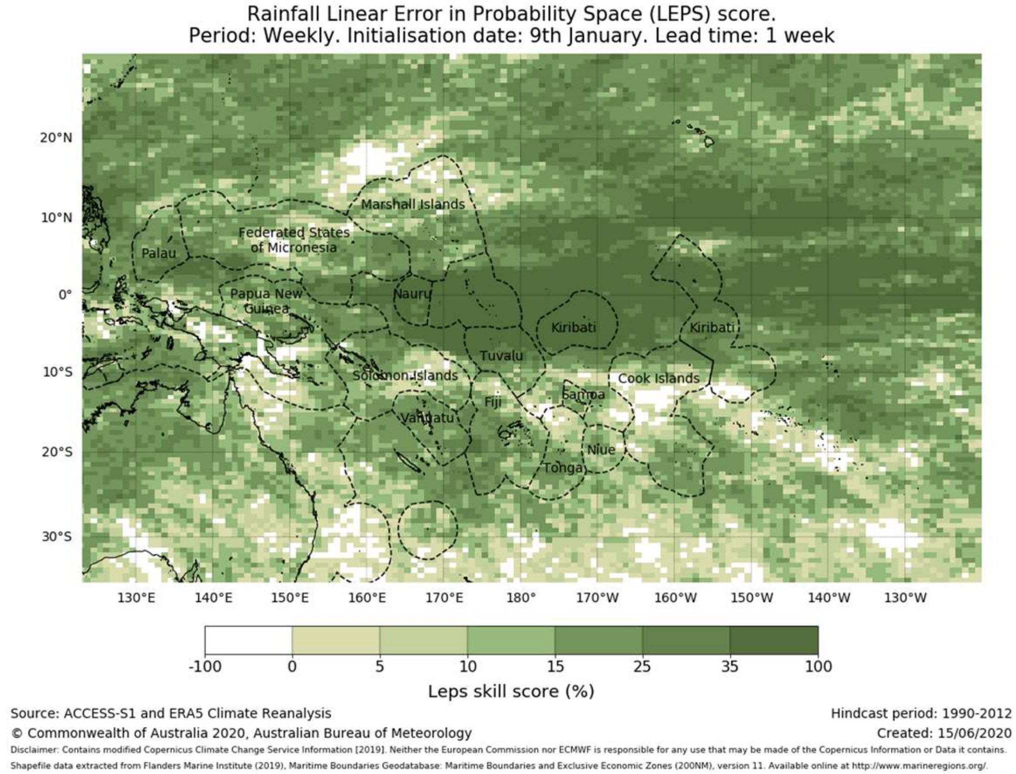
The LEPS percentage skill score
= SK = 100% * S / PF
= 100% * 0.00063/0.296
= 0.2%

Over the **history** of our forecasts, e.g. ACCESS hindcasts, the LEPS Skill Score =
100% * ∑ S / ∑ PF

Climate and Oceans Support Program in the Pacific

Converting LEPS to Words (OCOF)

Very Low: LEPS < 0.0
Low: LEPS of 0.0 to 5.0
Moderate: LEPS of 5.1 to 10.0
Good: LEPS of 10.1 to 15.0
High: LEPS of 15.1 to 25.0
Very High: LEPS of 25.1 to 35.0
Exceptional: LEPS > 35.0



What is "truth" when verifying a forecast?

The "**truth**" data that we use to verify forecasts generally comes from **observational data**. Such as:

- Rain gauge measurements
- Temperature observations
- Satellite-derived cloud cover
- Geopotential height analyses

It is difficult to know the exact truth because there are **errors** and **uncertainty** in the **observations** such as:

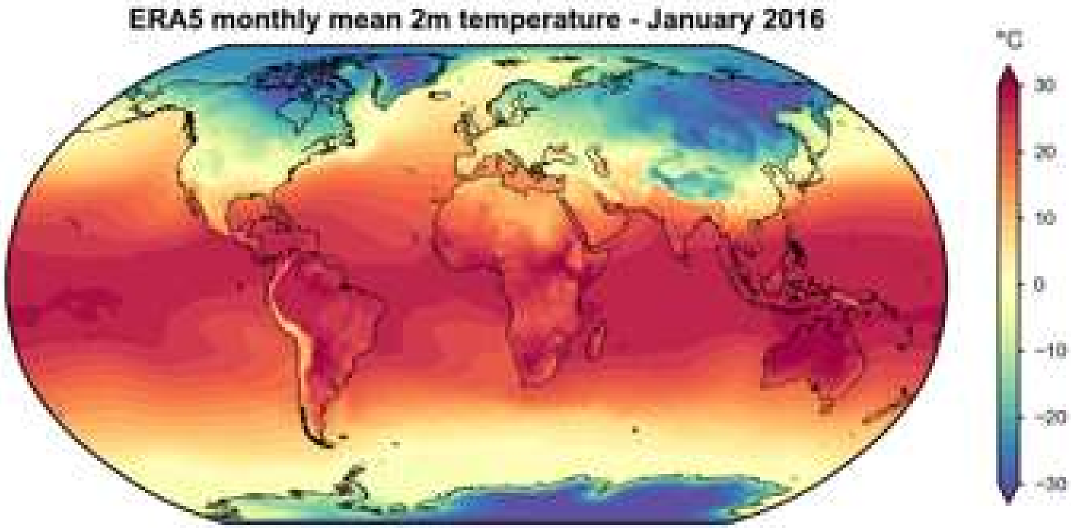
- Random and bias errors in the measurements
- Sampling errors
- Analysis error
- When data is altered to match the scale of the forecast

ERA5 reanalysis – what is it and why are we using it?

Station observations are the best, but in the **Pacific** there are **not enough for verification**.

Use gridded 'observational' data instead.

ECMWF uses its forecast models and data assimilation systems to '**reanalyse**' archived observations, creating global data sets describing the recent history of the atmosphere, land surface, and oceans.



ERA5, produced by C3S at ECMWF, is the current atmospheric reanalysis and is based on a 2016 version of the Integrated Forecasting System (IFS).

ERA5 is available hourly from 1979 to five days before present at a 0.25° x 0.25° resolution (30km grid). Atmospheric variables available on 37 pressure levels.

This can be used for verification.

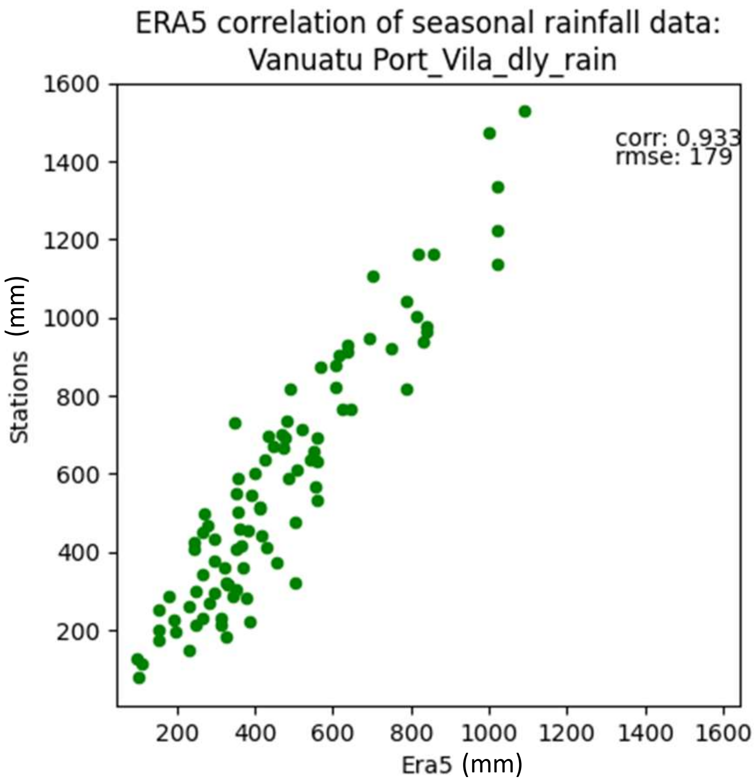
https://confluence.ecmwf.int/display/CKB/ERA5%3A+data+documentation

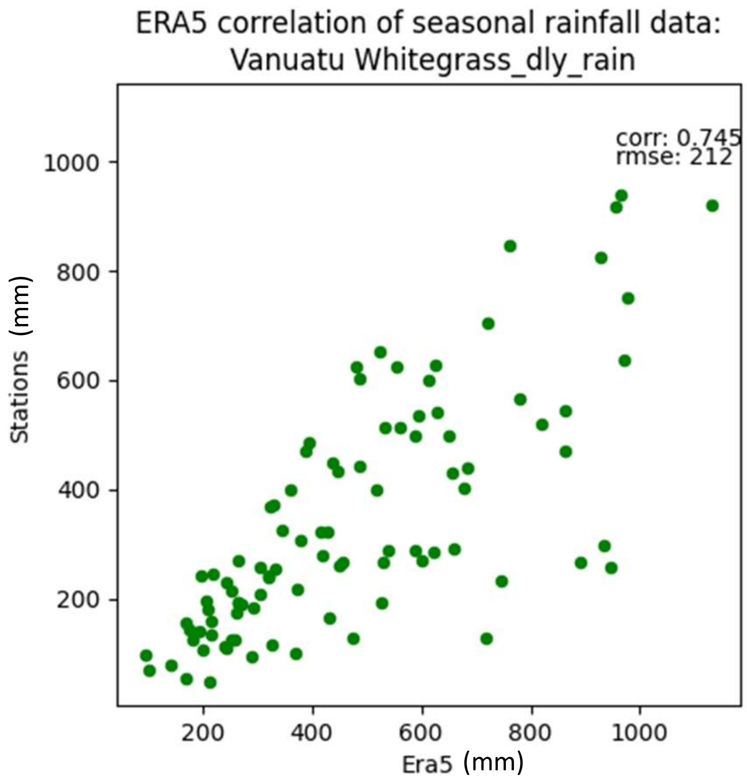
ERA5 compared to station observations

Climate and Oceans Support Program in the Pacific

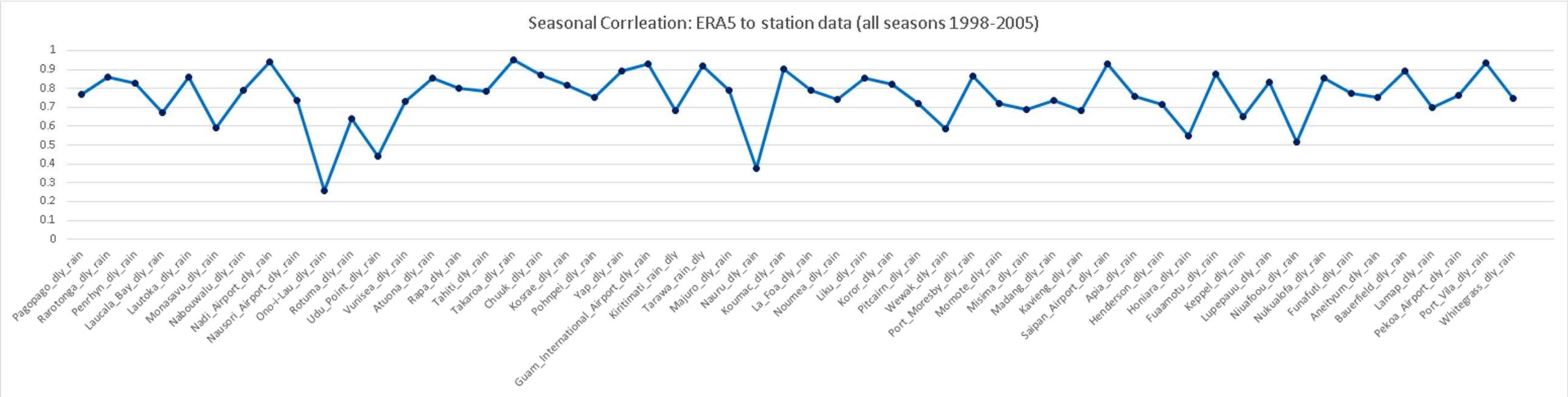
- ERA5 is a decent estimate of rainfall across the Pacific, but some regions are better than others
- Global data is available from 1979 up to 5 days ago covering all regions not just land based. This allows for hindcast and real-time model verification over the entire globe

Vanuatu example. All seasons (computed from daily data) 1998-2005

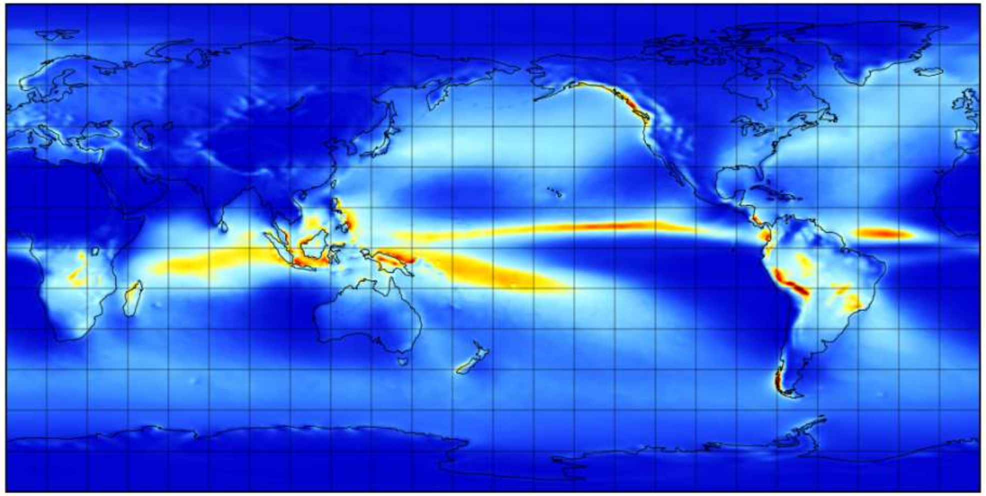




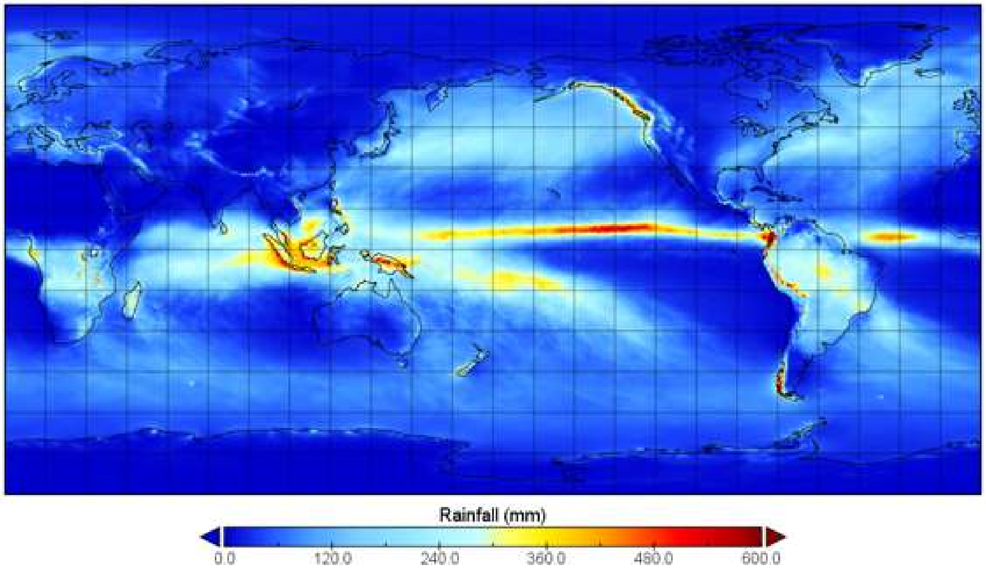
Seasonal correlation: ERA5 to station data



ACCESS-S and ERA5 climatology: December example



ACCESS-S1 hindcast climatology from 9th of November runs for lead time 1 (December).
11 ensemble members (model was run 11 times)
Climatology is the average rainfall from the 11 ensemble members for all 9th of November runs from 1990-2012



ERA5 reanalysis
IFS model run with extensive observations and satellite data from all December's 1990-2012.

ACCESS-S1 performs well compared to gridded observations

Skill summary

- Verification is important (we need to know if and where our model does a good job)
- Verifying forecasts is essential to improving future forecasts
- Seasonal climate forecasts require complex techniques to verify forecasts
- Linear Error in Probability Space (LEPS) used for ACCESS-S
- ERA5 reanalysis is used for verification, its not perfect but good enough