Why Do Forecast Trials Often Fail to Answer the Questions for which End-Users Need Answers: A Forecaster’s Point of View

Craig Collier

UVIG Forecasting Workshop, 2017

Ungraded

> We know there are some things we do not know. But there are also unknown unknowns, the ones we don’t know we don’t know.

- Donald Rumsfeld, Feb 12, 2002

Ungraded

The Trial Trilemma

Three priorities for trial setup

Fairness

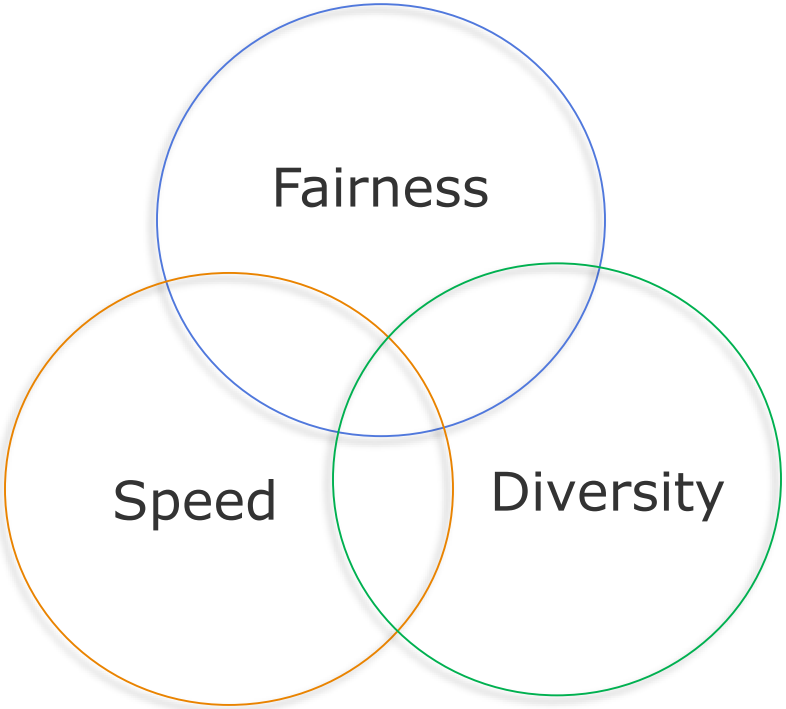
- Unbiased
- Standardized

Diversity

- Extendible
- Sufficient

Speed

- Ordered, with deadlines
- Limited
- Decision-driven



Ungraded

Questions We Want to Answer

Trials attempt to answer several important questions:

1. Which vendor will have the lowest error?
2. Which vendor's forecast is most correlated with actual generation?
3. Which vendor solution has the greatest range/applicability?
4. Which vendor offers the best balance of cost and performance?

Many others, but these are some of the most important

Ungraded

An Experiment

Let’s use real data to simulate a wind forecast trial (and proceeding 12-month performance period)

Experimental Design

- Three (3) independent model solutions to represent 3 independent, unique forecast vendors
- Models have no prior training data, and the same real-time data provided to each at exactly the same time every day during the trial period
- Trial period runs for one (1) month, randomly chosen.
- Forecasts will be provided for 3 actual sites, each separated by ~ 2300 km
- No expectation to predict outages, availability, or curtailments.
- Budget allows for only **one vendor** to get the contract, based on DA performance.

Ungraded

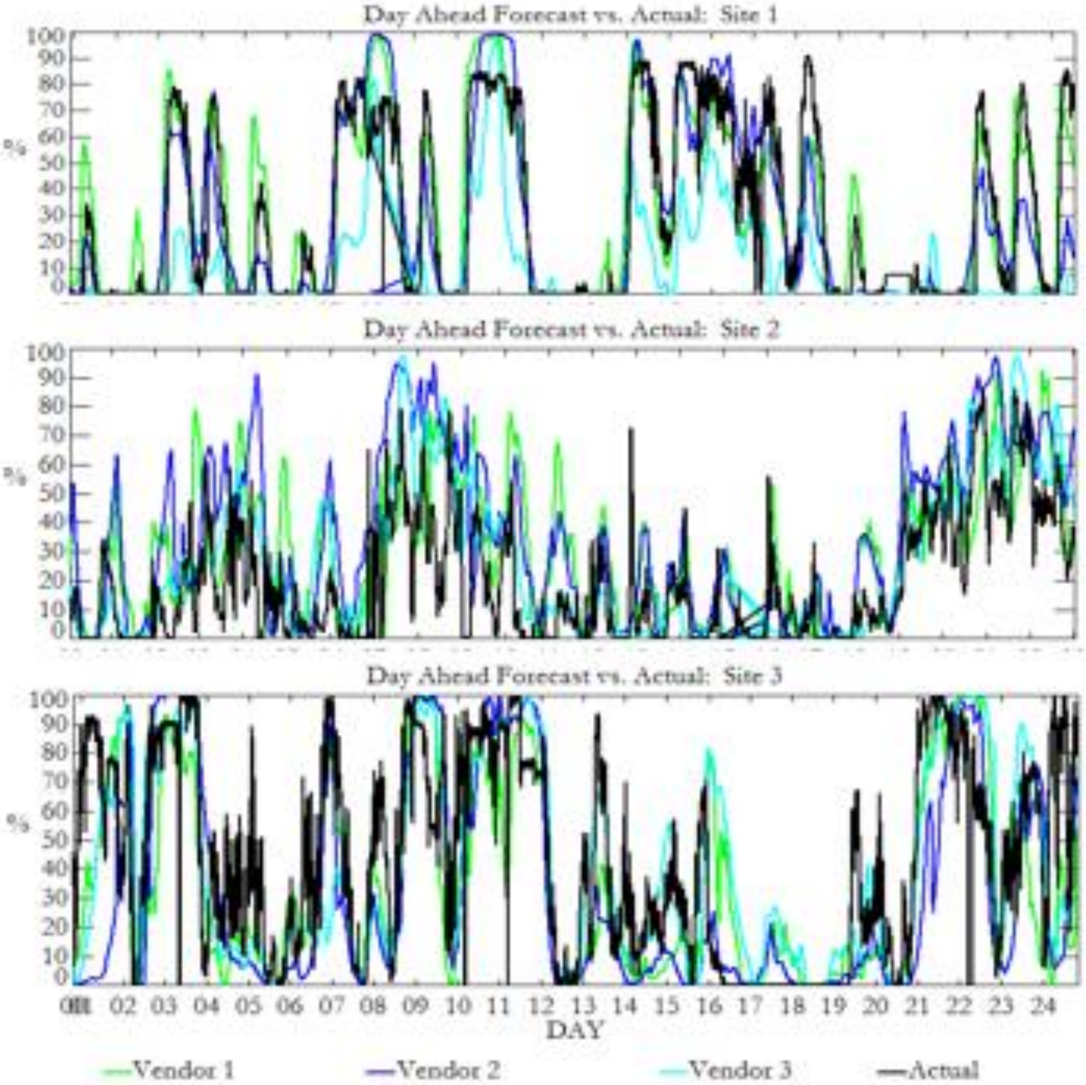
Diversity of Solution & Diversity of Site

Day-Ahead Forecasts

Trial sites meet requirements for diversity, sufficiency of challenge.

Trial site production unconstrained and reasonable.

Dispersion amongst the vendors – not always possible to achieve such spread.



Ungraded

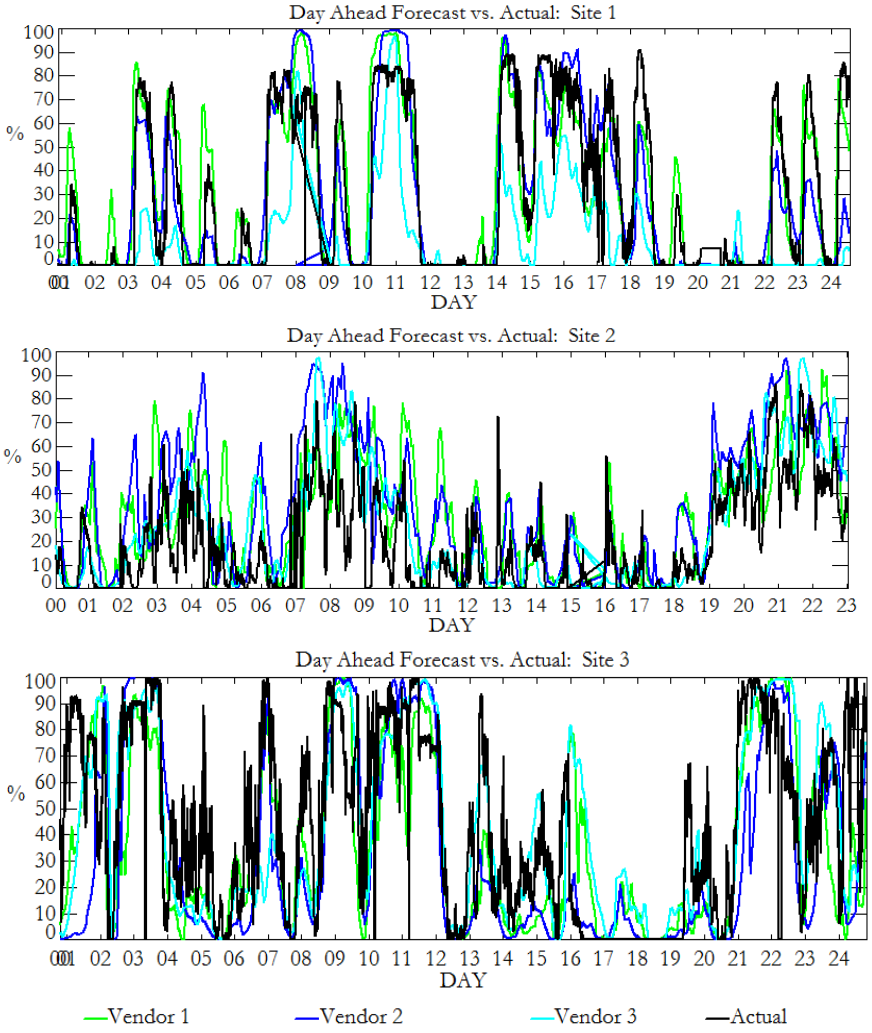
Trial Month:
Vendor Performance Relative to Average

Day-Ahead Results

rMAPE	V1 ●	V2 ●	V3 ○
Site 1	-3%	-2%	+6%
Site 2	-1%	0%	-1%
Site 3	0%	0%	-1%

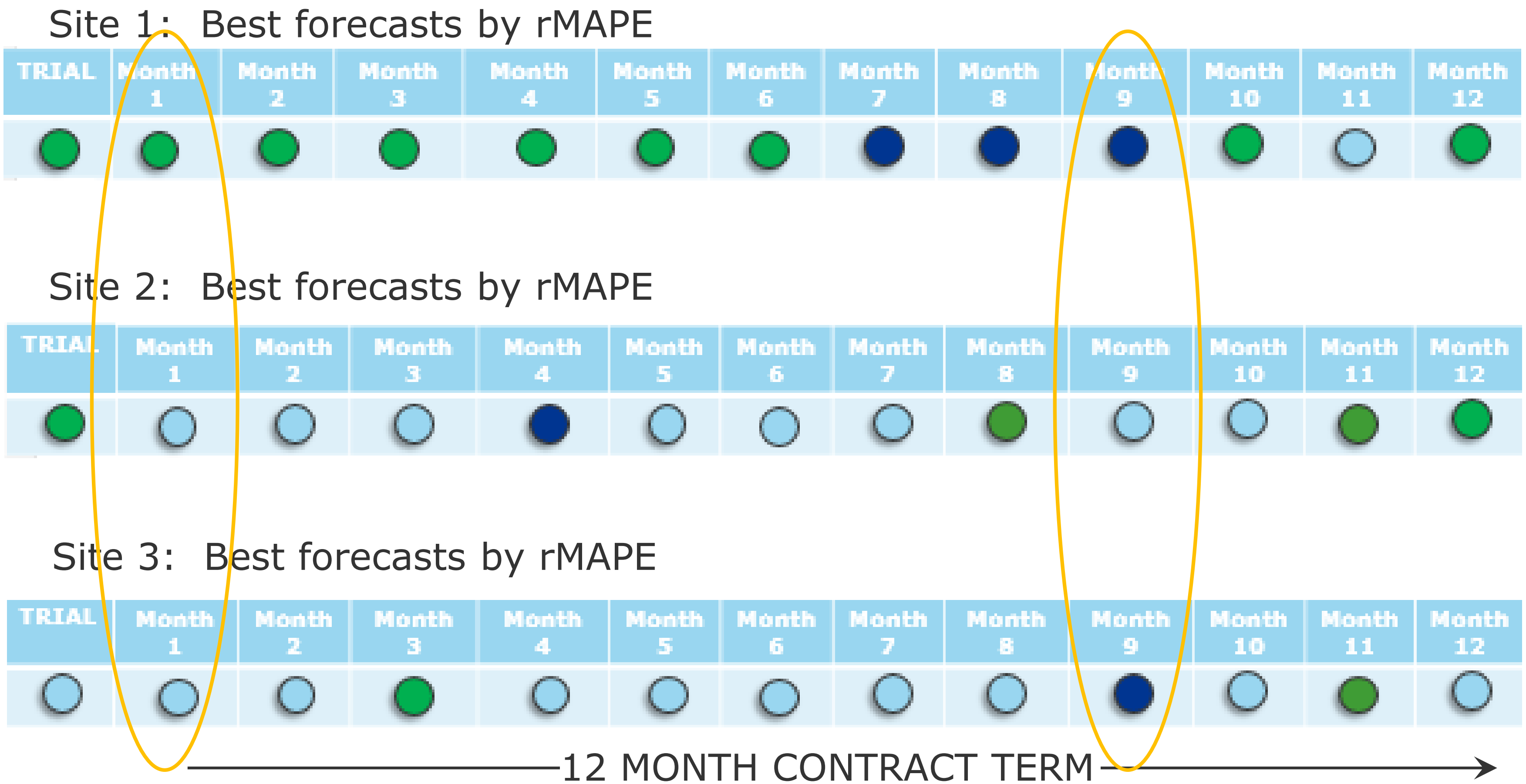
rCORR	V1 ●	V2 ●	V3 ○
Site 1	0.7	0.6	0.5
Site 2	0.5	0.6	0.5
Site 3	0.6	0.5	0.7

While Vendors 1 & 3 are nearly a toss-up, Vendor 3 disappoints on site 1 more than Vendor 1 disappoints on site 3.



Ungraded

Reliability: Is Performance Sustained?



Site 1: Best forecasts by rMAPE

TRIAL	Month 1	Month 2	Month 3	Month 4	Month 5	Month 6	Month 7	Month 8	Month 9	Month 10	Month 11	Month 12
🟢	🟢	🟢	🟢	🟢	🟢	🟢	🔵	🔵	🔵	🟢	⚪	🟢

Site 2: Best forecasts by rMAPE

TRIAL	Month 1	Month 2	Month 3	Month 4	Month 5	Month 6	Month 7	Month 8	Month 9	Month 10	Month 11	Month 12
🟢	⚪	⚪	⚪	🔵	⚪	⚪	⚪	🟢	⚪	⚪	🟢	🟢

Site 3: Best forecasts by rMAPE

TRIAL	Month 1	Month 2	Month 3	Month 4	Month 5	Month 6	Month 7	Month 8	Month 9	Month 10	Month 11	Month 12
⚪	⚪	⚪	🟢	⚪	⚪	⚪	⚪	⚪	🔵	⚪	🟢	⚪

12 MONTH CONTRACT TERM

🟢 Vendor 1

🔵 Vendor 2

⚪ Vendor 3

Ungraded

Does Timing Matter?

- In trial month, Vendor 1 exhibited lowest error and greatest range, **BUT**…
  - Delayed **1-2** months: Vendor 3 scores highest for MAPE & Range
  - Delayed **9** months: Vendor 2 scores highest for MAPE & Range

- For this portfolio, the trial selection repeatable 40% of the time

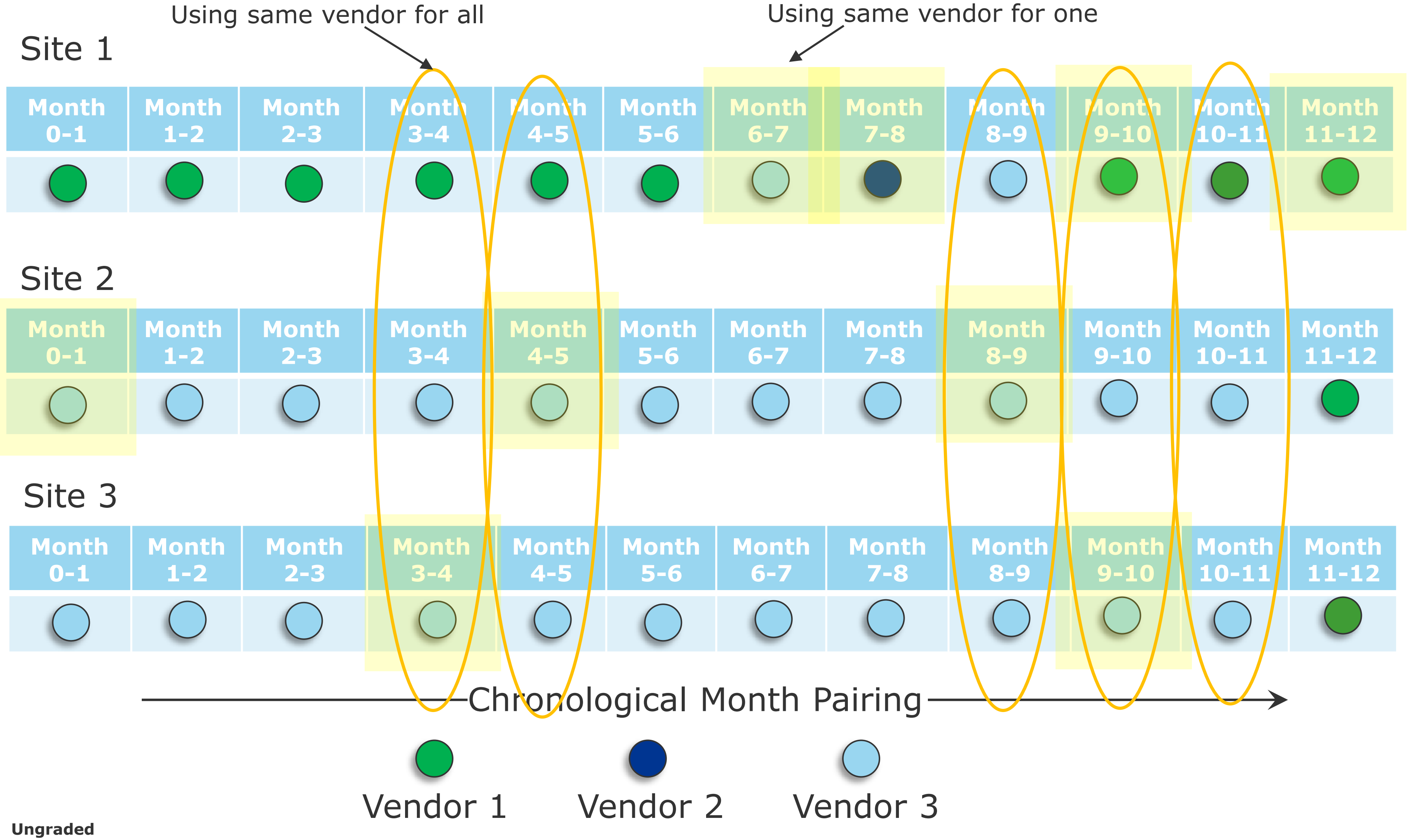
- For a single site, the trial selection repeatable 75-80% of the time

- In a 30-day trial, reliability of the solution over a 12-month term is difficult to measure

- Selecting more than 1 vendor increases the probability of reliability

Ungraded

Effect of Trial Duration



Using same vendor for all

Using same vendor for one

Site 1

Month 0-1 | Month 1-2 | Month 2-3 | Month 3-4 | Month 4-5 | Month 5-6 | Month 6-7 | Month 7-8 | Month 8-9 | Month 9-10 | Month 10-11 | Month 11-12

Site 2

Month 0-1 | Month 1-2 | Month 2-3 | Month 3-4 | Month 4-5 | Month 5-6 | Month 6-7 | Month 7-8 | Month 8-9 | Month 9-10 | Month 10-11 | Month 11-12

Site 3

Month 0-1 | Month 1-2 | Month 2-3 | Month 3-4 | Month 4-5 | Month 5-6 | Month 6-7 | Month 7-8 | Month 8-9 | Month 9-10 | Month 10-11 | Month 11-12

Chronological Month Pairing

Vendor 1

Vendor 2

Vendor 3

Sensitivity to Trial Duration

- An extra 30 days changes the outcome for a single portfolio selection. **Vendor 3** would have been the likely selection.

- For this portfolio, the trial selection was repeatable 92% of the time with an extra 30 days.

- For the individual site, the trial selection was repeatable at least 75% of the time.

- Solution reliability is enhanced by doubling duration but is ***not guaranteed***.

- Need to strongly consider the **costs to the vendor** for doubling duration.

- What are the accuracy-related costs for settling on one vendor vs. the costs of integrating two?

Hard and Soft Characteristics

Traditionally, forecast trials are based on hard characteristics: availability of forecast MW, Met, Uncertainty, update frequency, granularity, MAPE, Bias.

Soft Characteristics comprise the features, services, and support surrounding the *hard* offering

Alerts : automated or manual indicators of extreme events

Meteorological expertise: situational awareness from atmospheric scientists. We need to answer:

- Why is the forecast behaving this way?
- Can the forecast be believed?
- What are the drivers?

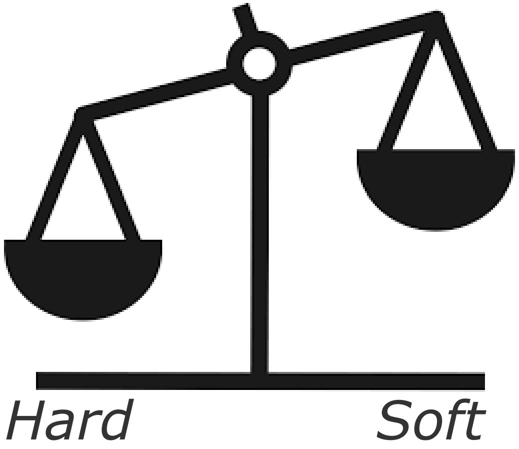
Customization: Helping the user integrate the forecast into decision support mechanism

Support: Reachability and accessibility of the vendor

Ungraded

The Value of Soft Characteristics

Hard characteristics always get more weight than soft characteristics – as it should be



Should they be appraised in a trial?

How would we value soft characteristics empirically? Can they be indexed?

P (Operational Support) = P (Not Reasonable U Not Available)

P (Custom Support) = P (Knowledge Gap U Capability Gap)

Ungraded

The Irony of Soft Characteristics

P (Operational Support) = P (Not Reasonable U Not Available) → **0 (in trial)**

P (Custom Support) = P (Knowledge Gap U Capability Gap) → **0 (in trial)**

In reality,

0% < P (Not Available) < 1% P (Not Reasonable) > 1%

→ P (Operational Support) ≠ 0

A solution evolves:

→ P (Custom Support) ≠ 0

Trials measure neither the probabilities or adequacy of response

Conclusions

- Forecast trials are not answering the questions for which users need answers due to the inherent constraints of trial design. A trial is a ***sample,*** primarily focused on a single metric (and cost).

- Probability of solution reliability can be enhanced but never guaranteed. For a total portfolio / single vendor approach, probability is enhanced by trial duration, but for single site/single vendor, 30 days likely sufficient.

- Diversity of solution mitigates the uncertainty of solution reliability – but user-integration cost should be balanced against opportunity cost of single provider.

- Operational and custom support are not measured in trial – but probabilities of occurrence in operation are not zero and should never be considered zero.

Ungraded

Thank You

Craig Collier, Ph.D.

Section Head, Forecasting

craig.collier@dnvgl.com
(858) 836–3370, ext. 118

www.dnvgl.com

SAFER, SMARTER, GREENER

Ungraded