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
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
The Trial Trilemma

Three priorities for trial setup

Fairness
- Unbiased
- Standardized

Diversity
- Extendible
- Sufficient

Speed
- Ordered, with deadlines
- Limited
- Decision-driven
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
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.
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.
### Day-Ahead Results

<table>
<thead>
<tr>
<th></th>
<th>V1</th>
<th>V2</th>
<th>V3</th>
</tr>
</thead>
<tbody>
<tr>
<td>Site 1</td>
<td>-3%</td>
<td>-2%</td>
<td>+6%</td>
</tr>
<tr>
<td>Site 2</td>
<td>-1%</td>
<td>0%</td>
<td>-1%</td>
</tr>
<tr>
<td>Site 3</td>
<td>0%</td>
<td>0%</td>
<td>-1%</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>V1</th>
<th>V2</th>
<th>V3</th>
</tr>
</thead>
<tbody>
<tr>
<td>Site 1</td>
<td>0.7</td>
<td>0.6</td>
<td>0.5</td>
</tr>
<tr>
<td>Site 2</td>
<td>0.5</td>
<td>0.6</td>
<td>0.5</td>
</tr>
<tr>
<td>Site 3</td>
<td>0.6</td>
<td>0.5</td>
<td>0.7</td>
</tr>
</tbody>
</table>

While Vendors 1 & 3 are nearly a toss-up, Vendor 3 disappoints on site 1 more than Vendor 1 disappoints on site 3.
Reliability: Is Performance Sustained?

Site 1: Best forecasts by rMAPE

Site 2: Best forecasts by rMAPE

Site 3: Best forecasts by rMAPE

12 MONTH CONTRACT TERM

Vendor 1  Vendor 2  Vendor 3
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
# Effect of Trial Duration

## Using same vendor for all

<table>
<thead>
<tr>
<th>Month 0-1</th>
<th>Month 1-2</th>
<th>Month 2-3</th>
<th>Month 3-4</th>
<th>Month 4-5</th>
<th>Month 5-6</th>
<th>Month 6-7</th>
<th>Month 7-8</th>
<th>Month 8-9</th>
<th>Month 9-10</th>
<th>Month 10-11</th>
<th>Month 11-12</th>
</tr>
</thead>
<tbody>
<tr>
<td>Site 1</td>
<td>Site 2</td>
<td>Site 3</td>
<td>Site 1</td>
<td>Site 2</td>
<td>Site 3</td>
<td>Site 1</td>
<td>Site 2</td>
<td>Site 3</td>
<td>Site 1</td>
<td>Site 2</td>
<td>Site 3</td>
</tr>
</tbody>
</table>

## Using same vendor for one

<table>
<thead>
<tr>
<th>Month 0-1</th>
<th>Month 1-2</th>
<th>Month 2-3</th>
<th>Month 3-4</th>
<th>Month 4-5</th>
<th>Month 5-6</th>
<th>Month 6-7</th>
<th>Month 7-8</th>
<th>Month 8-9</th>
<th>Month 9-10</th>
<th>Month 10-11</th>
<th>Month 11-12</th>
</tr>
</thead>
<tbody>
<tr>
<td>Site 1</td>
<td>Site 2</td>
<td>Site 3</td>
<td>Site 1</td>
<td>Site 2</td>
<td>Site 3</td>
<td>Site 1</td>
<td>Site 2</td>
<td>Site 3</td>
<td>Site 1</td>
<td>Site 2</td>
<td>Site 3</td>
</tr>
</tbody>
</table>

## Chronological Month Pairing

- **Vendor 1**
- **Vendor 2**
- **Vendor 3**

**Ungraded**
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
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?

\[
P \text{ (Operational Support)} = P \text{ (Not Reasonable OR Not Available)}
\]

\[
P \text{ (Custom Support)} = P \text{ (Knowledge Gap OR Capability Gap)}
\]
The Irony of Soft Characteristics

\[ P(\text{Operational Support}) = P(\text{Not Reasonable } \cup \text{ Not Available}) \rightarrow 0 \text{ (in trial)} \]

\[ P(\text{Custom Support}) = P(\text{Knowledge Gap } \cup \text{ Capability Gap}) \rightarrow 0 \text{ (in trial)} \]

In reality,

\[ 0\% < P(\text{Not Available}) < 1\% \quad P(\text{Not Reasonable}) > 1\% \]

\[ \rightarrow P(\text{Operational Support}) \neq 0 \]

A solution evolves:

\[ \rightarrow P(\text{Custom Support}) \neq 0 \]

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.
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