New Directions in Statistical Post-Processing

Tom Hamill

NOAA ESRL Physical Sciences Division

Boulder, Colorado USA

tom.hamill@noaa.gov

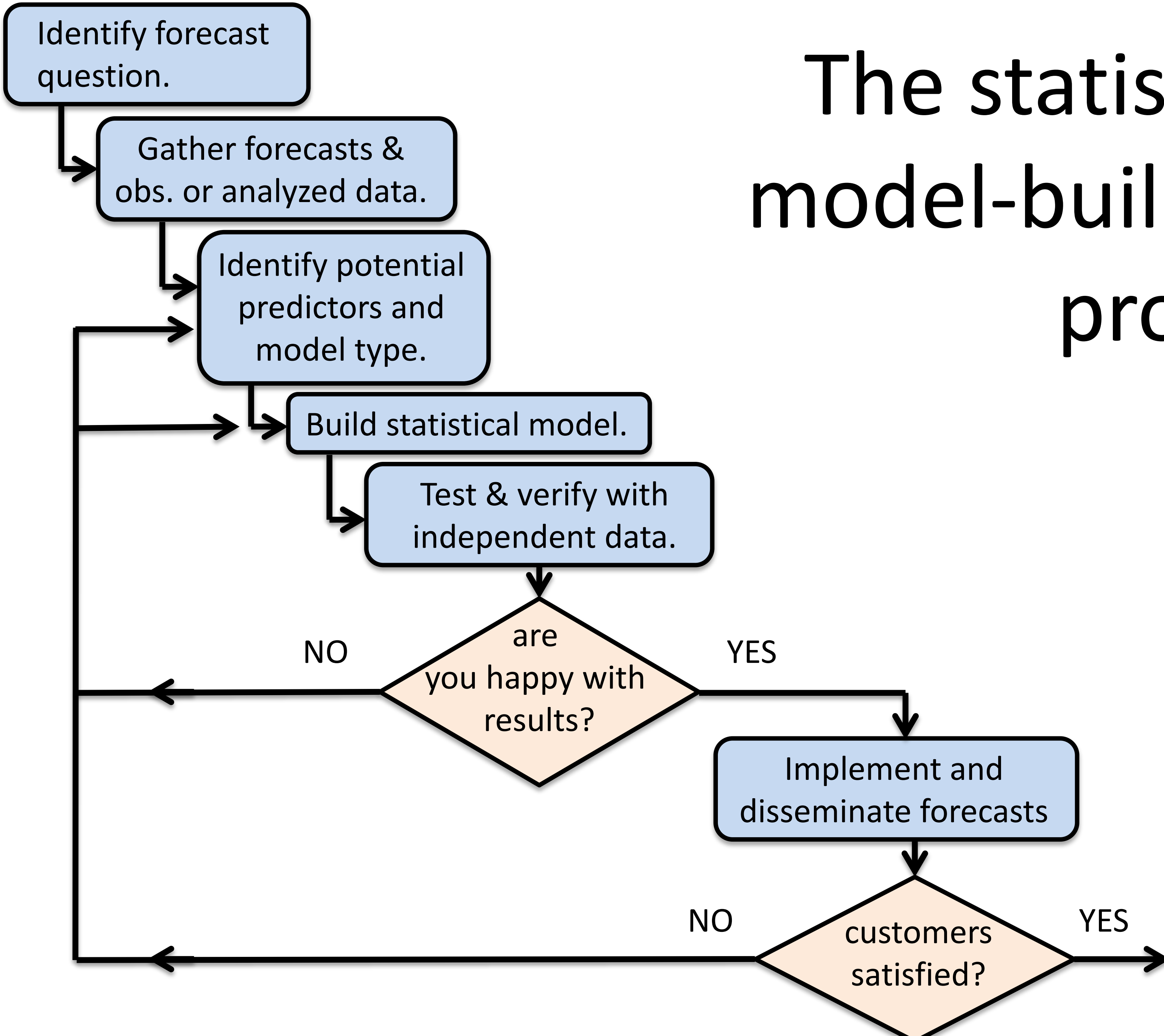
Acknowledgments: Michael Scheuerer, CIRES

An invited presentation to Glahn Symposium, 2015 AMS Annual Meeting

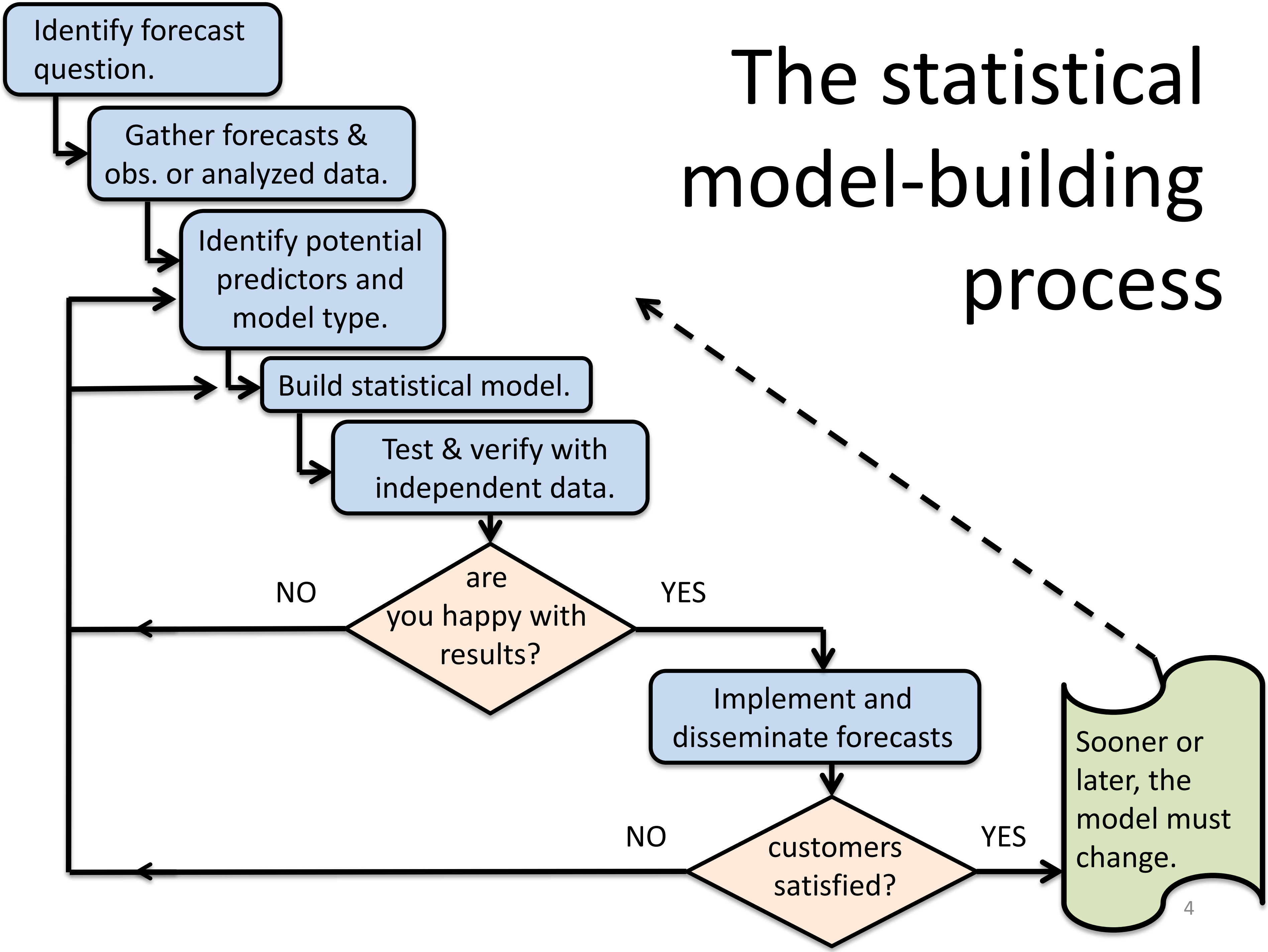
We mean many things by "statistical post-processing"

- Distribution fitting.
- Perfect-prog methods.
- Physically based statistical models (e.g. DeMaria's LGEM).
- Model output statistics (MOS; thanks Bob Glahn)
  - Implicitly, many methods, not just multiple linear regression.
  - Develop predictive relationships between past observed and forecast.
  - From this, estimate probability distribution of observed given today's forecast.
- Etc.

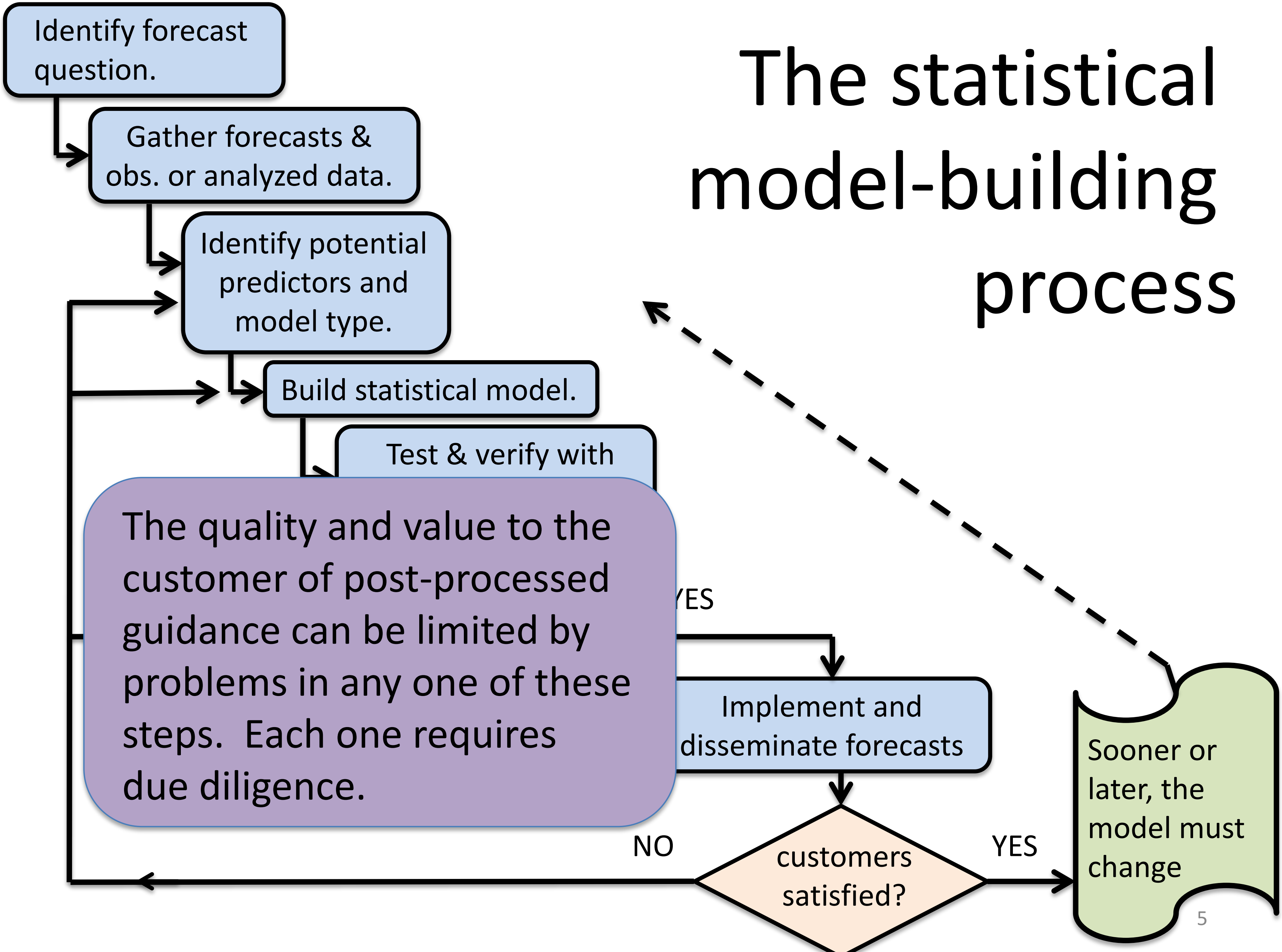
The statistical model-building process



The statistical model-building process

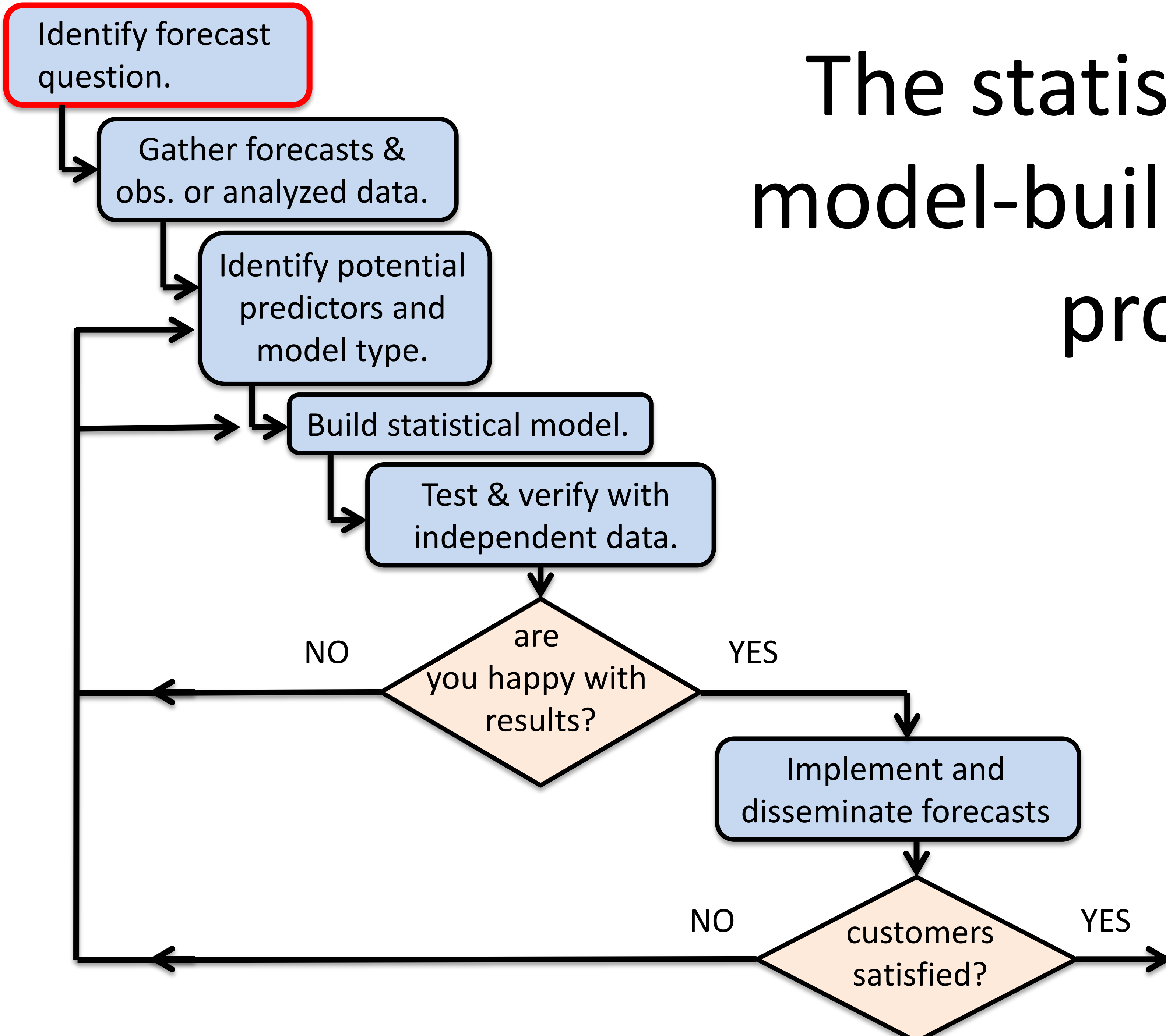


The statistical model-building process

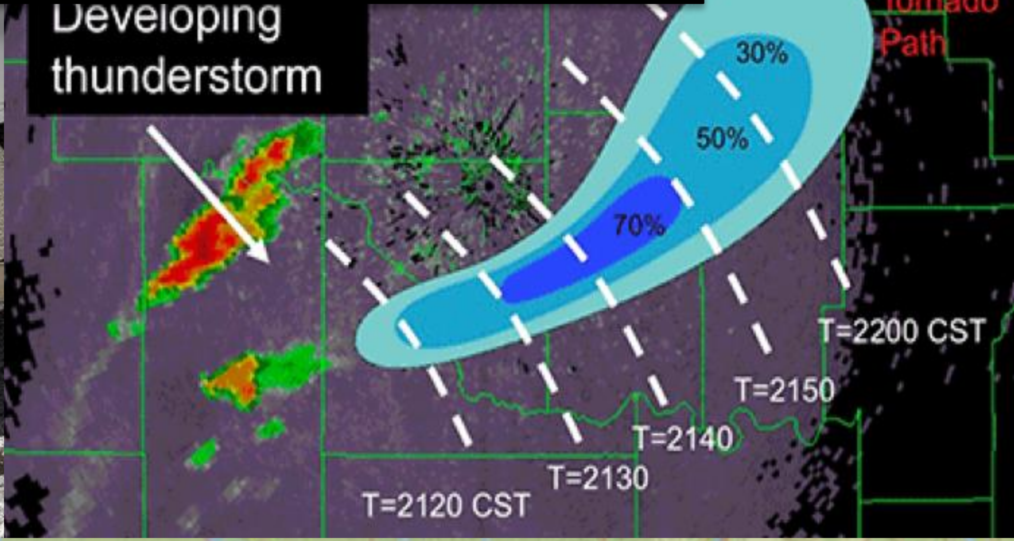


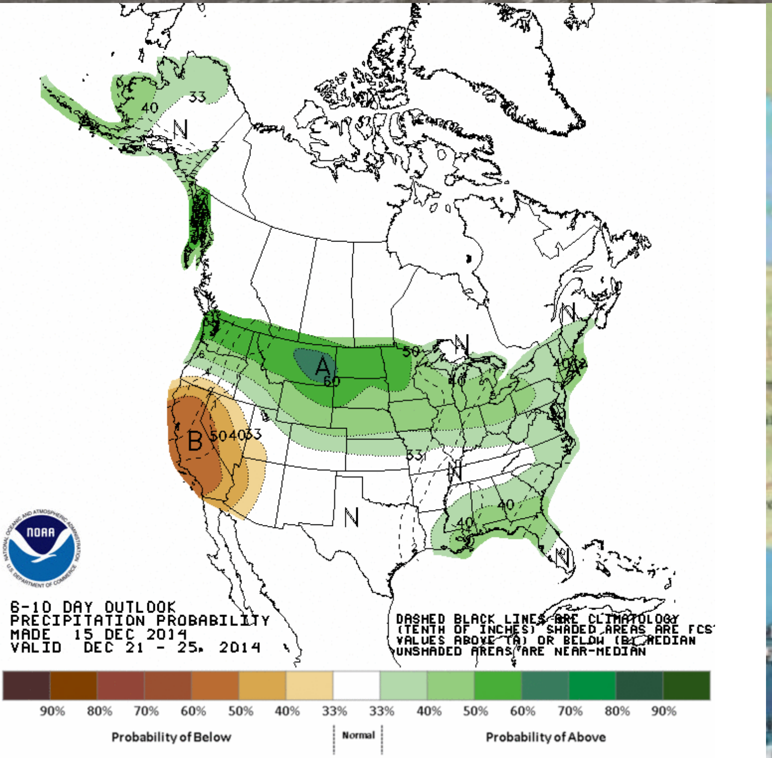
The quality and value to the customer of post-processed guidance can be limited by problems in any one of these steps. Each one requires due diligence.

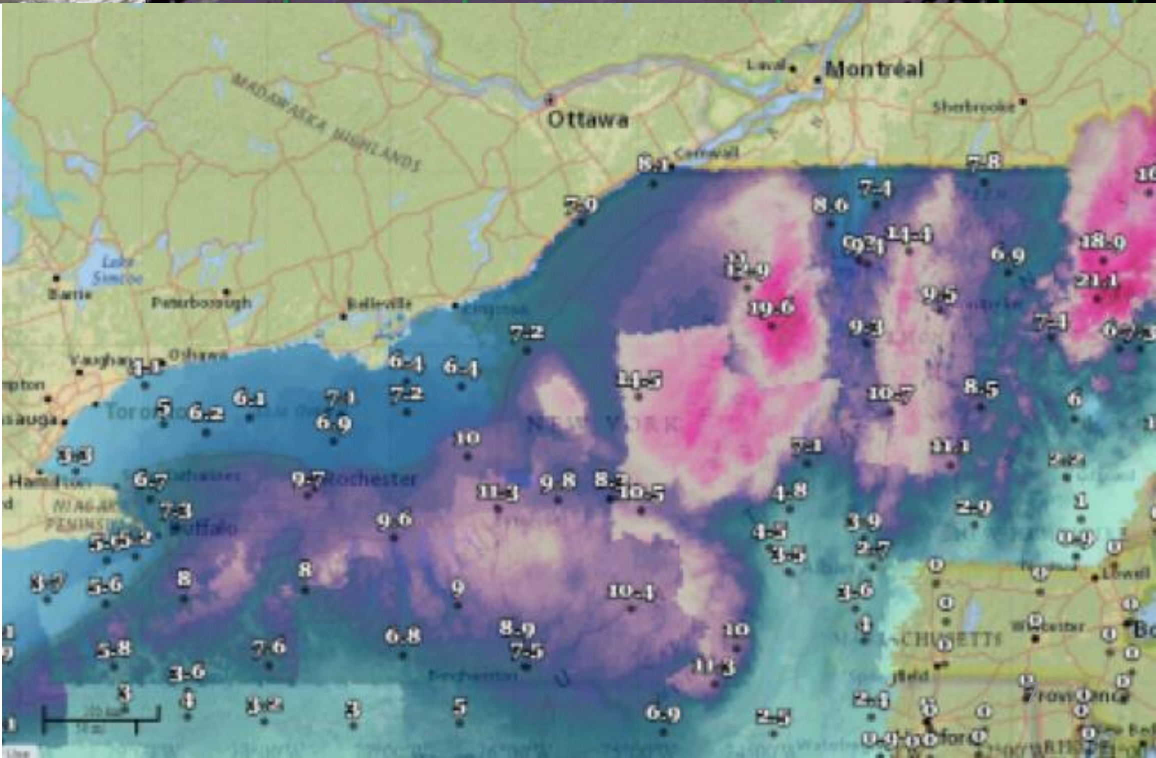
The statistical model-building process



Identifying the forecast question







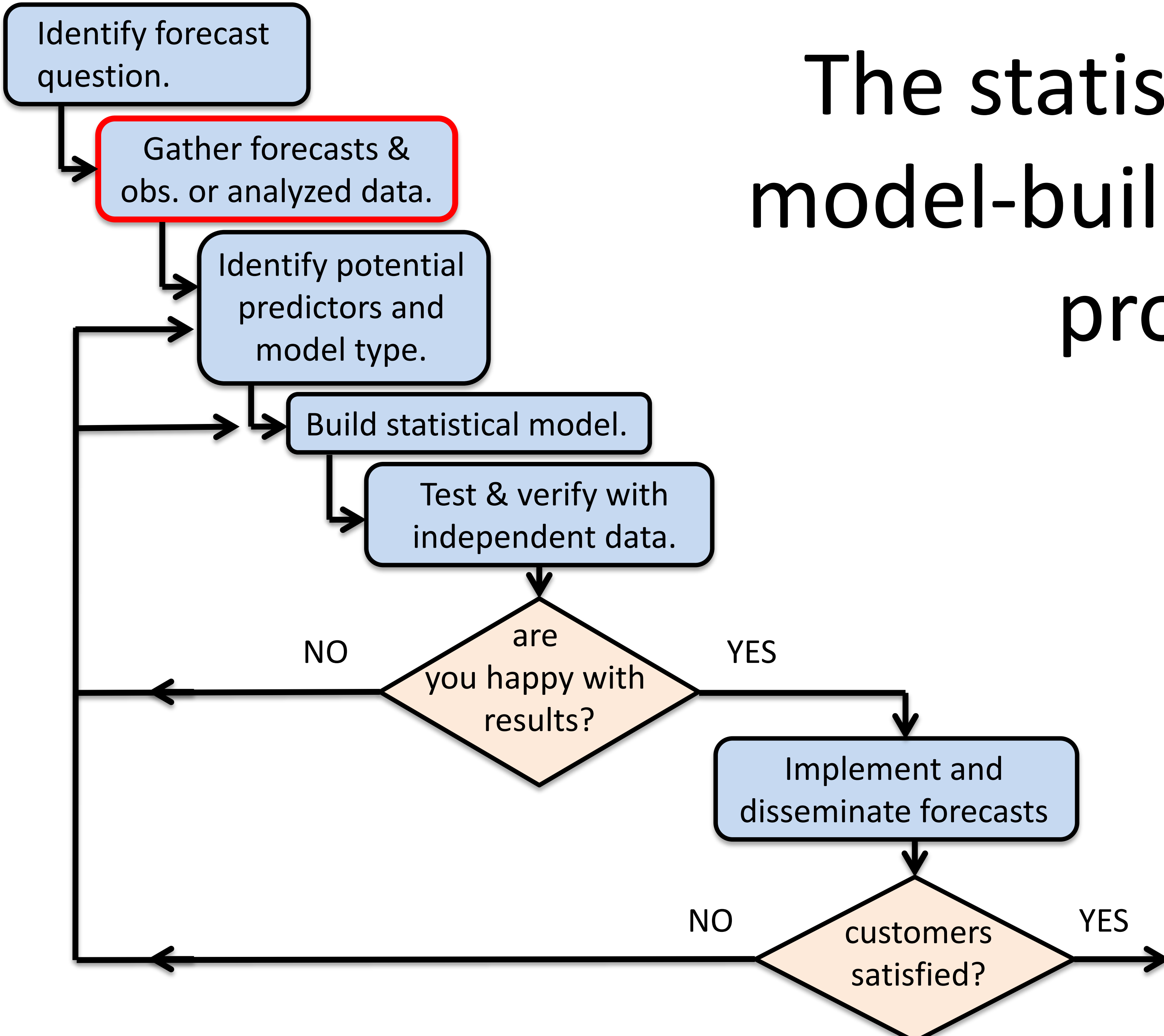
What are customers asking for?

*Increasingly, post-processed guidance for weather related to **high-impact** events.*

- Applications
  - Precipitation amount and type, and drought.
  - Cloud amount, type, ceiling, visibility, insolation (for solar energy).
  - Aviation: Icing, turbulence, winds en route, thunderstorm areal coverage.
  - Ship routing, wave height.
  - Wind, gustiness.
  - Wind power and its "ramps."
  - Tropical cyclogenesis probabilities, TC intensity, location.
  - Tornadoes and severe weather.
  - Temperature, humidity.
- Characteristics:
  - High-resolution (spatial & temporal).
  - Low-error deterministic and reliable & sharp probabilistic.
  - Probabilities of extremes.
  - Time scales: nowcast to decadal-centennial.
  - Spatial and temporal correlation structures (multi-variate).
- And so forth.

Post-processing resources for development and maintenance are limited. How to choose?

The statistical model-building process



Gathering forecasts, observations, & analyzed data.

Higher-resolution models, more models, run more frequently

Improved assimilation methods

Improved physics

Frequent model updates & bug fixes

More ensemble members

High-quality reanalyses for initialization, statistical model development, verification

Retrospective forecasts

More stable models

HPC funds to disk space for rapid access to past forecasts

Can we resolve this tension?

NCEP/EMC plans for evolution of model implementation process

(from 2014 Production Suite Review)

- **GFS**: implementations yearly, with 2-3 years of reanalyses and reforecasts.
- **GEFS**: implementations every other year, with reanalysis &reforecast from ~ 1999-present.
- **CFS**: implementations every 4th year, with modern-era reanalysis reforecast.

We are grateful to NCEP.

Now we need spiffy new methods worthy of this rich data.

Data set details to iron out.

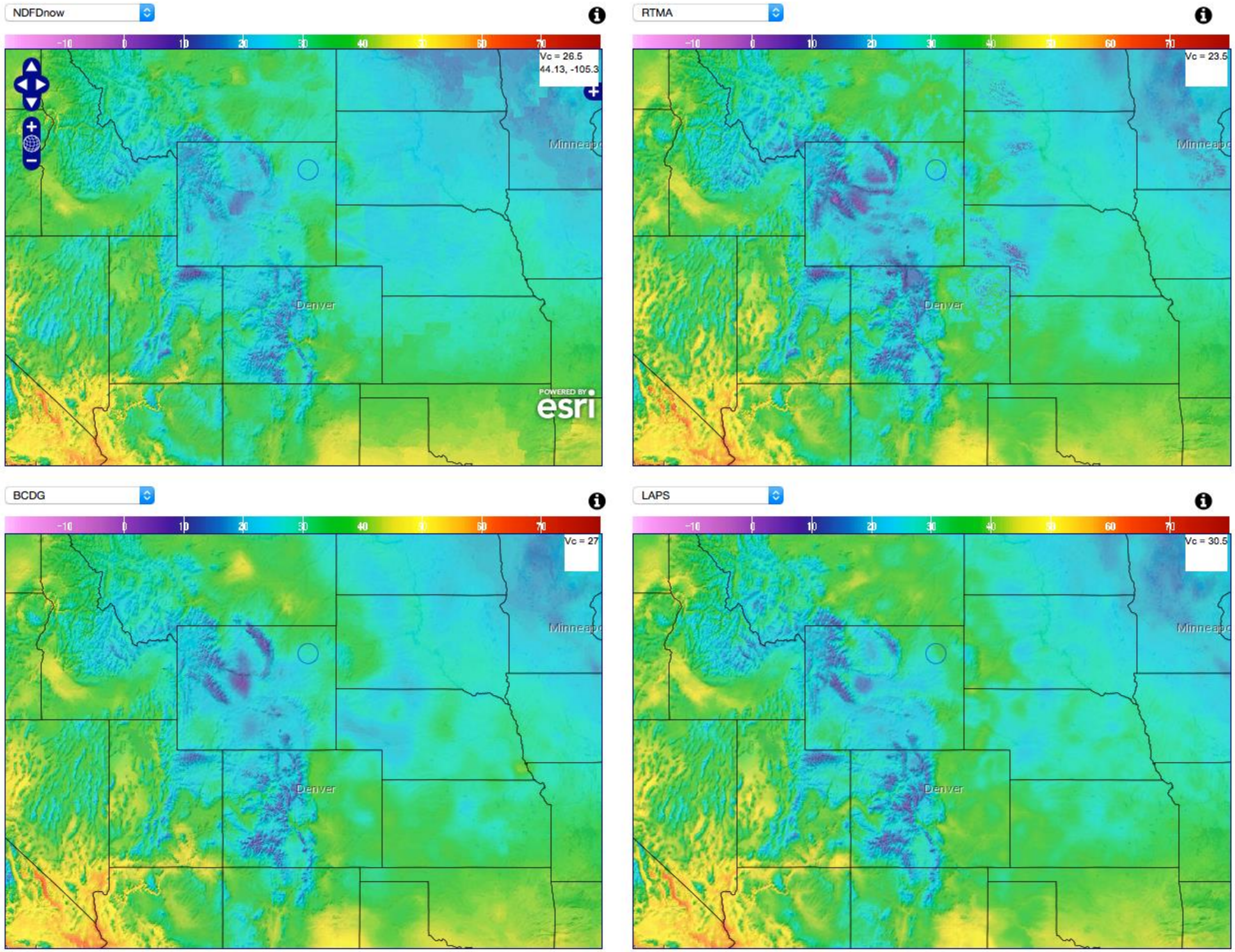
- Reanalyses.
  - *Regular* production for reforecast initialization; same model, same resolution, same assimilation methodology.
  - High-res., high-quality surface reanalyses for training, verification.
- Reforecast
  - # members/cycle? # cycles/day? Frequency? How far into the past? Structure satisfactory to all while being computationally tractable? [See NOAA white paper for a start]
  - ~ homogeneous forecast errors and bias over reforecast period.
- Plentiful, *non-proprietary* observations (e.g., precipitation type, severe weather, stream flow, wind power)
- Robust supporting infrastructure.
  - HPC for reanalysis, reforecast.
  - Large amount of rapid-access disk space.
  - Computer cycles for post-processing model development.
  - Bandwidth for dissemination of high-res. probabilistic products.

Retrospective analyses: RTMA/URMA

- NWS's ~ 2.5 - 3 km mesoscale hourly surface analysis
  - covers N America, AK, HI, PR, Guam
  - temp, dewpoint, winds, visibility.
  - used for verification, training in prominent "National Blend" project.
- Implementation soon: 3-km HRRR and 4-km NAM blend for first-guess forecasts.
- Improving, but still concerns about analysis quality, esp. in mountainous terrain.
- Will need RTMA run in past to cover the same period as reforecasts.

http://www.mdl.nws.noaa.gov/~blend/blender.prototype.php

Noticeable differences between mesoscale analyses



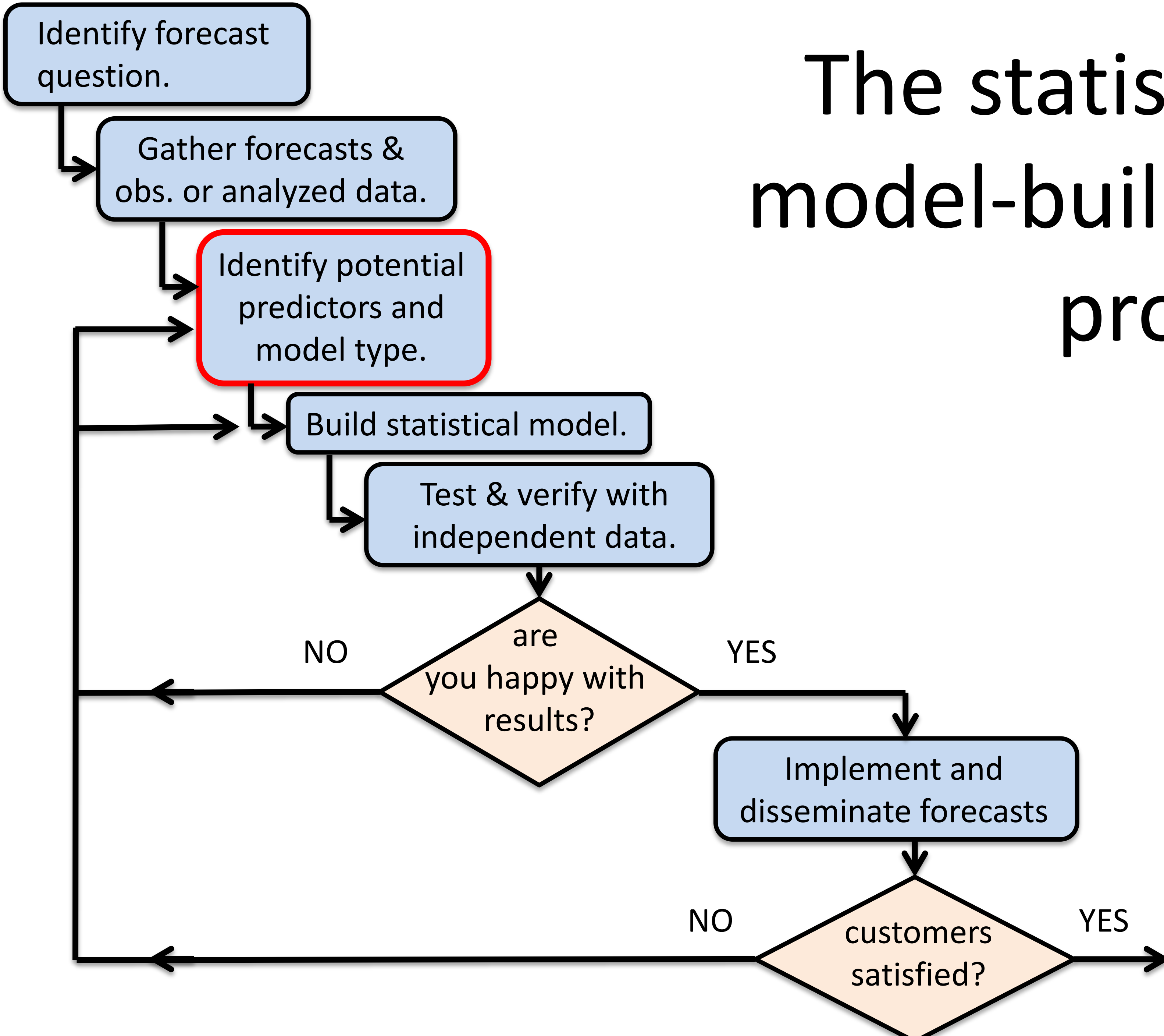
http://www.mdl.nws.noaa.gov/~blend/blender.prototype.php (internal NOAA)

Reforecast sample size: how many?

Amount	Post-processing application
a few are sufficient	Short-range forecasts of surface temperature, dew point
	Short-range wind forecasts
	Forecasts of light precipitation events
	Wind-power "ramp" events and wind-error spatial correlations
	Extended-range temperature and precipitation, temperature extremes
a very large number needed	Forecasts of heavy precipitation events
	Tornado forecasts
	Space-time correlation structure of hydrologic forecast errors

There is no one optimal reforecast configuration for all applications.

The statistical model-building process

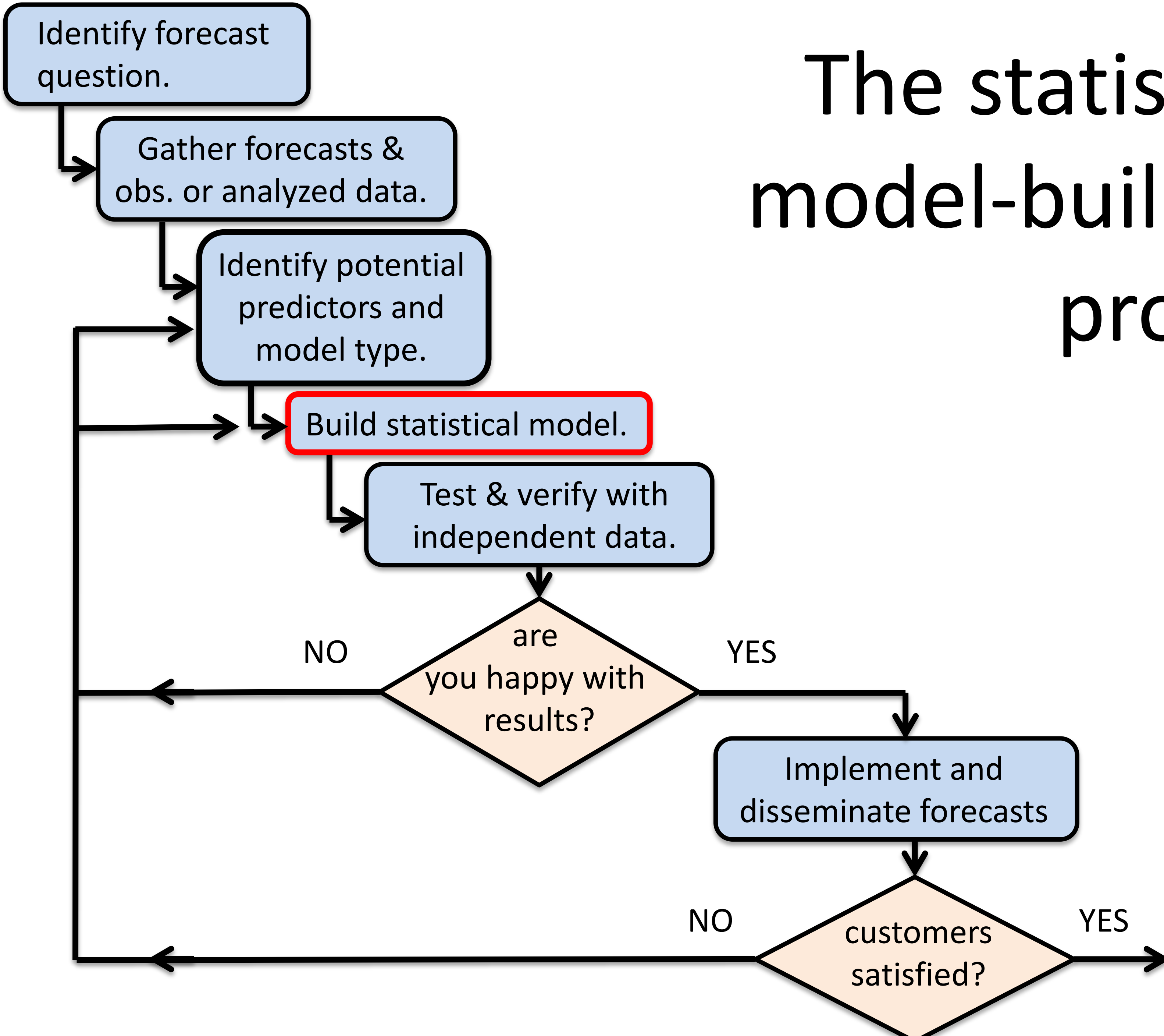


This step can be time consuming.

We need the exploratory data analysis and fast modeling tools of science fiction.

“Majority Report”

The statistical model-building process



Building a high-quality statistical model

- Old problems are new problems
  - *“ bias-variance tradeoff ”*
  - *“ extrapolating the regression ”*
  - *“ curse of dimensionality ”*
- More modular, reusable software.

Building a high-quality statistical model

- Old problems are new problems
  - “ *bias-variance tradeoff* ”
  - “ *extrapolating the regression* ”
  - “ *curse of dimensionality* ”
- More modular, reusable software.

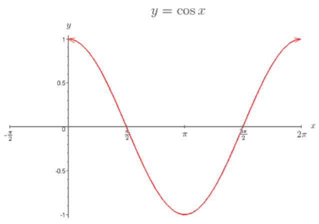
"Bias-variance tradeoff"

Example: let's demo NCEP/EMC's "decaying average" filter used in NAEFS for estimating bias with simulated data

Generate daily time series of truth, simulated observations (100x), and simulated forecasts (100x) under the condition of seasonally varying bias:

Truth T = 0.0 (always)

Bias: The true (but unknown) forecast bias for julian day t: $B_t = \cos(2\pi t/365)$



Analyzed for day t = truth + random obs error:

$x_t^A = T + e_t^A$, where $e_t^A \sim N(0, 1/3)$, iid each day.

Simulated biased forecasts for day t generated w. auto-correlated error via Markov Chain:

$x_t^f - B_t = k \bullet (x_{t-1}^f - B_{t-1}) + e_t^f$, where $e_t^f \sim N(0,1)$, iid each day; $k = 0.5$

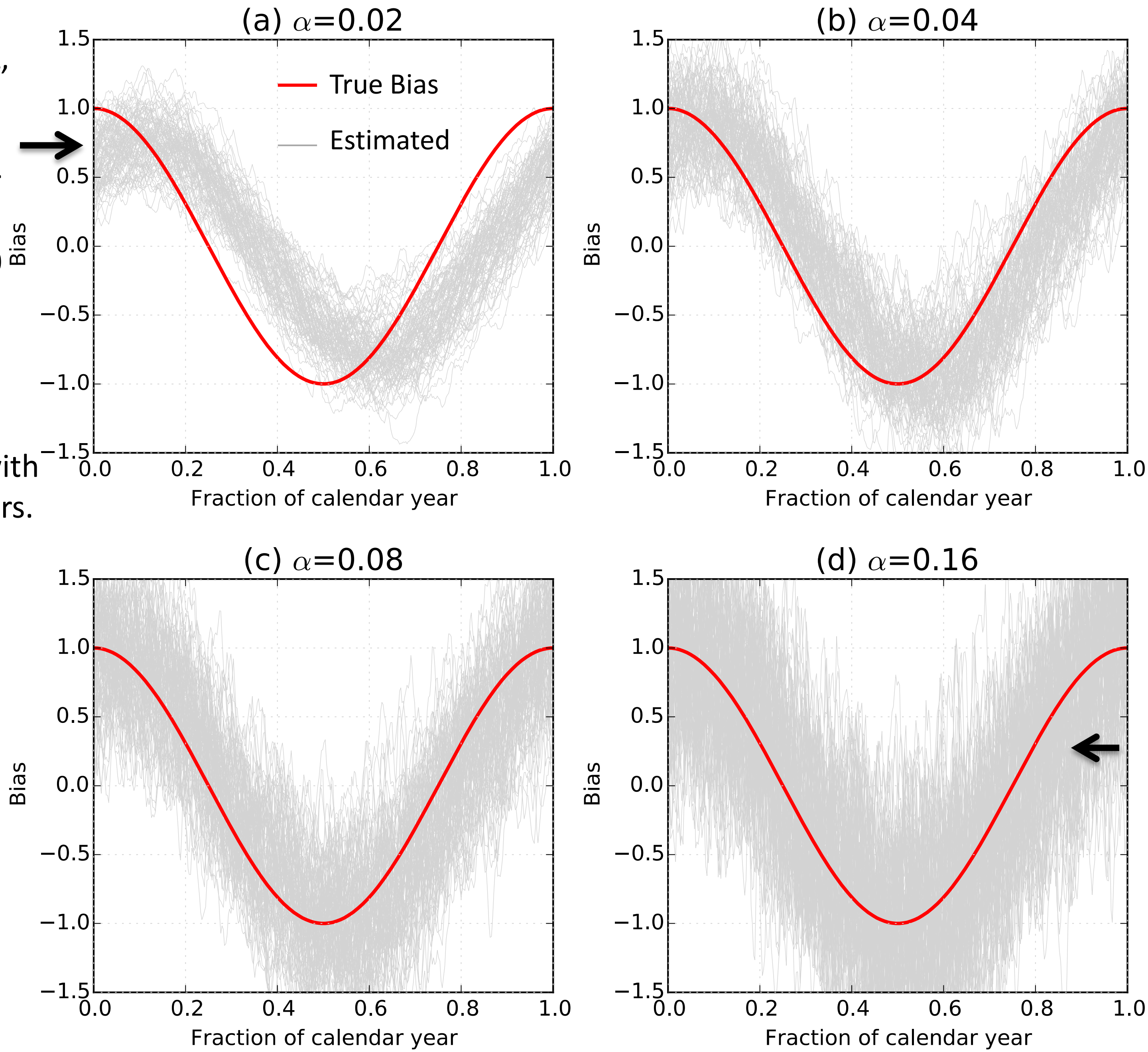
EMC's decaying-average bias correction: bias estimate B_t is

$$B_t = (1-\alpha) \bullet B_{t-1} + \alpha \bullet (x_t^f - x_t^A)$$

α is a user-defined parameter that indicates how much to weight most recent bias estimate; large α akin to overfitting in regression analysis.

Bias-Variance Tradeoff for Decaying Average Filter

With small α, the bias estimates have smaller variance over the 100 replications but lag the true bias. Akin to regression with few predictors.



With large α, bias estimates have large variance but don't much lag the true bias. Akin to regression with many predictors.

(plots shown after 60-day spinup)

Minimizing the bias-variance tradeoff

- Sometimes, change form of model and/or choice of predictors:

  Forecast = a + b•fcst + c•cos(2πt/365) + d*sin(2πt/365)

- Increase the training sample size.
  - Overall increase (reforecasts).
  - Selective increase*: estimate some parameters with local data, others with regional or global data.
  - Example: precipitation forecast adjustment.
    - Parameters related to terrain-related bias: local data.
    - Parameters related to ubiquitous drizzle over-forecast: regional data.

(* Stimulated by Scheuerer, & König, 2014: Gridded, locally calibrated, probabilistic temperature forecasts based on ensemble model output statistics. *QJRMS*, **140**, 2582-2590.

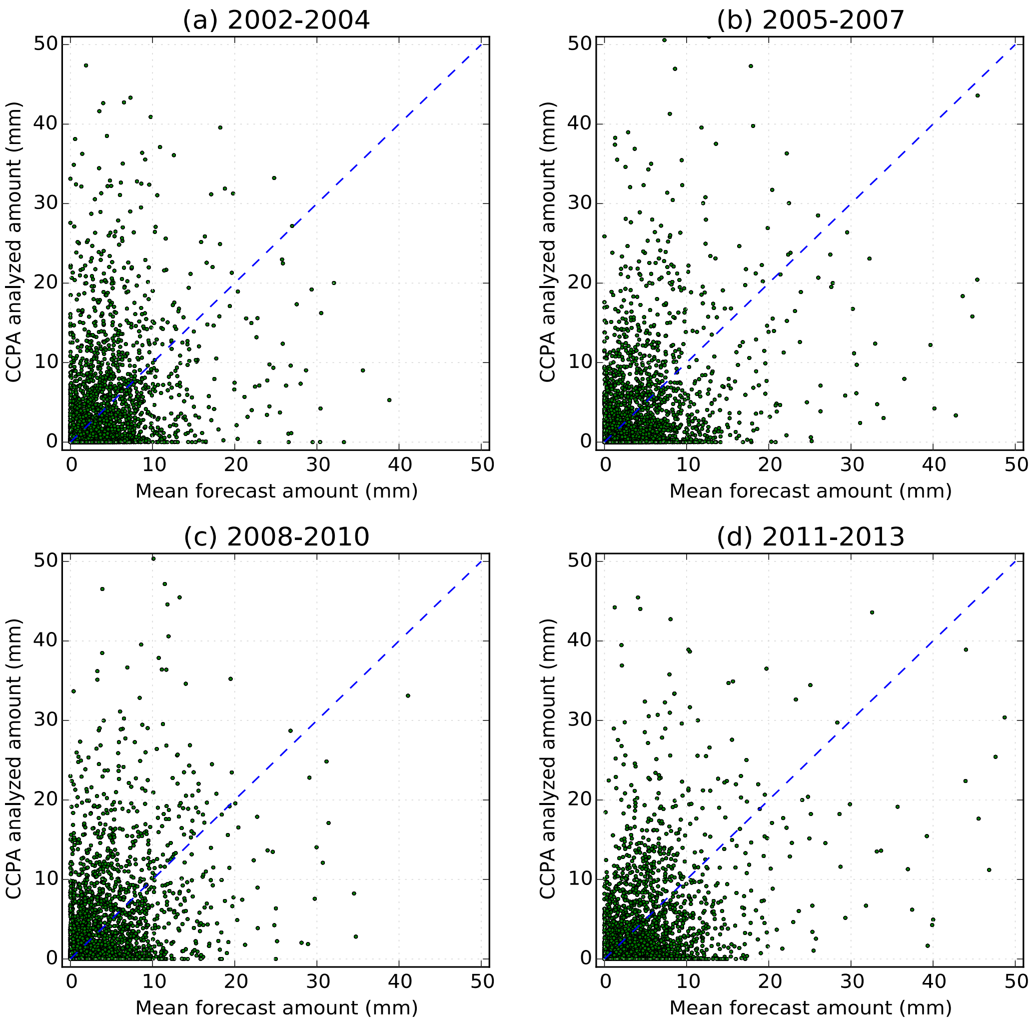
Building a high-quality statistical model

- Old problems are new problems
  - “ *bias-variance tradeoff* ”
  - “ *extrapolating the regression* ”
  - “ *curse of dimensionality* ”
- More modular, reusable software.

Extrapolating the regression

(that is, getting accurate predictions at and beyond the fringes of training data)

Ithaca, NY 2002-2013 Jun-Jul-Aug Precipitation, GEFS mean and CCPA analyzed



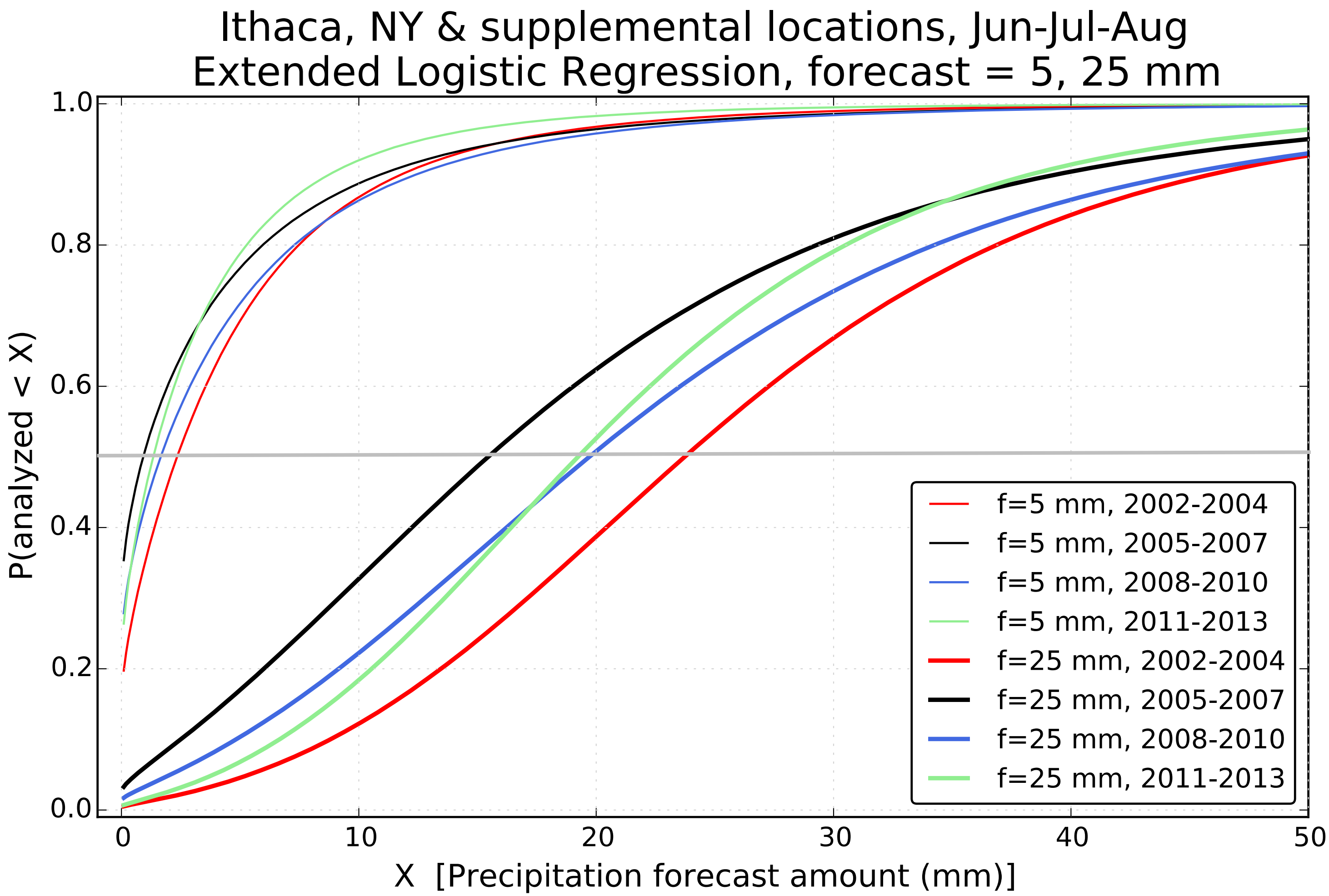
Data from grid point over Ithaca, NY, and 19 “supplemental” locations with similar climatologies, terrain features.

Split GEFS reforecast and analyzed precipitation data over the 2002-2013 period into 4 batches.

Perform Dan Wilks’ “extended logistic regression” (ELR) on each (power-transformed) batch.

The GEFS 36-48 h ensemble-mean forecast is the sole predictor, and the method produces PDFs or CDFs of predicted precipitation amount.

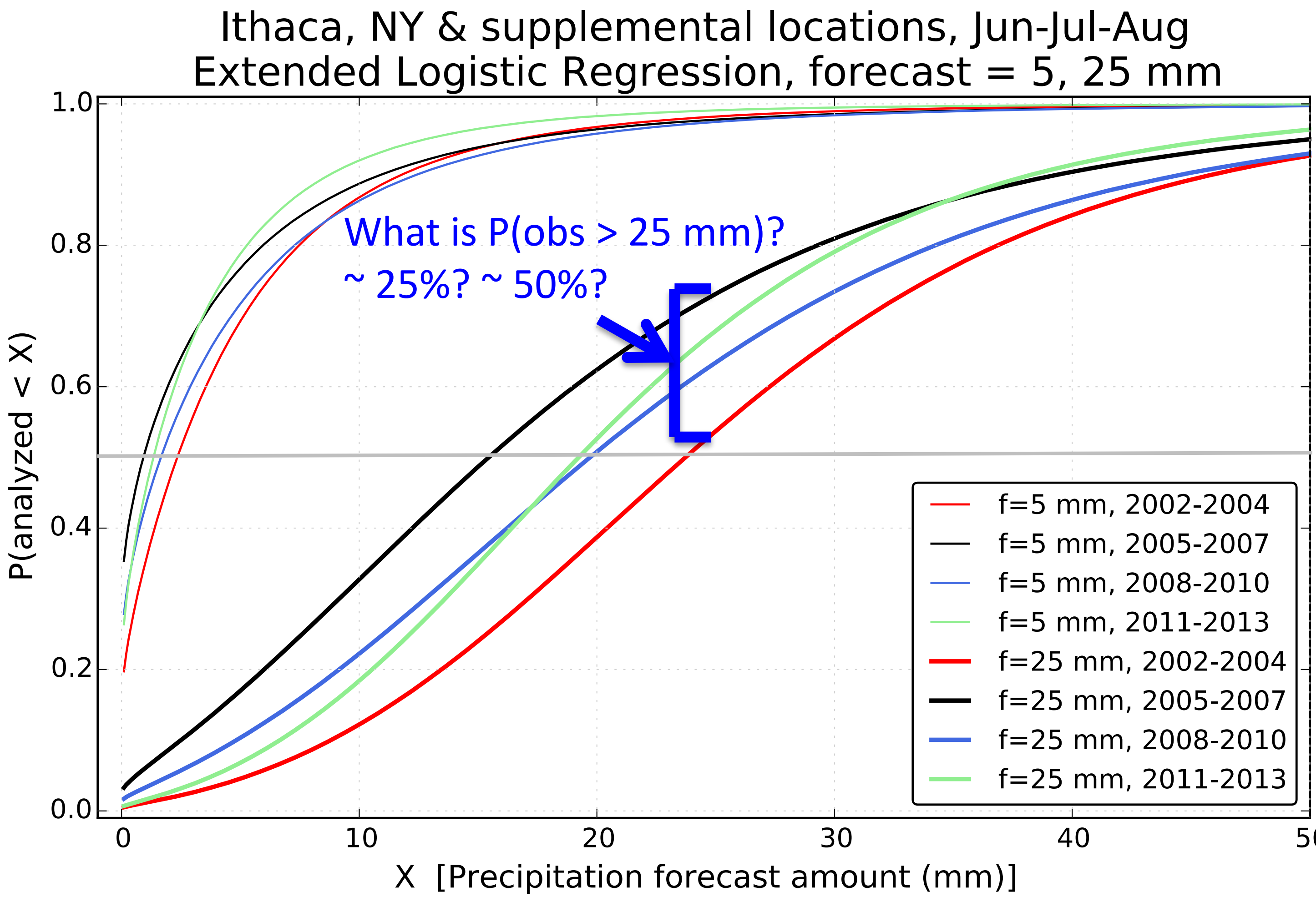
Predictive CDFs for the four batches
(given GEFS mean forecasts of 5 mm, 25 mm)



Despite the use of supplemental training data, there is great predictive uncertainty from ELR amongst the four batches when the forecast = 25 mm.

Predictive CDFs for the four batches

(given GEFS mean forecasts of 5 mm, 25 mm)



Despite the use of supplemental training data, there is great predictive uncertainty from ELR amongst the four batches when the forecast = 25 mm.

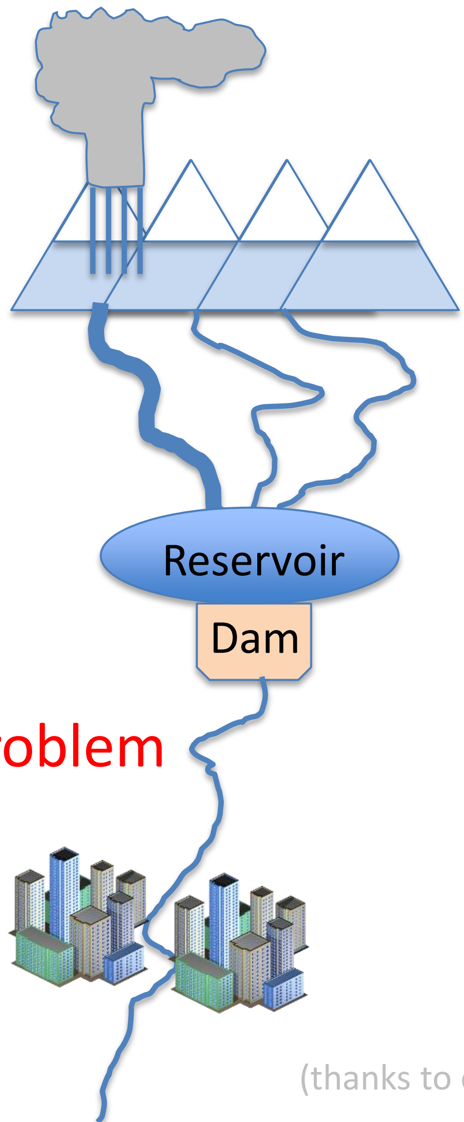
Ameliorating predictive model uncertainty when forecasts are extreme

- Some post-processing methods may be more sensitive than others, so test, test, test.
- Again, increase sample size.
  - Data from supplemental locations.
  - Reforecasts and analyses spanning decades.
- ID additional predictors with correlations to predictand.

Building a high-quality statistical model

- Old problems are new problems
  - “ *bias-variance tradeoff* ”
  - “ *extrapolating the regression* ”
  - “ *curse of dimensionality* ”
- More modular, reusable software.

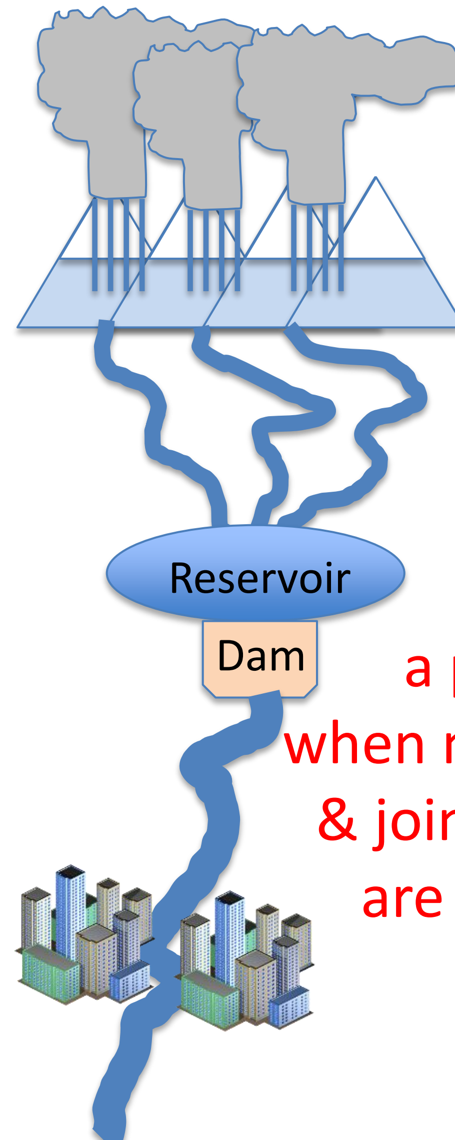
The curse of dimensionality: a motivation



No problem

Hydrologists want to know not only the intensity of rainfall, but whether or not that intense rainfall will fall simultaneously in many nearby sub-basins.

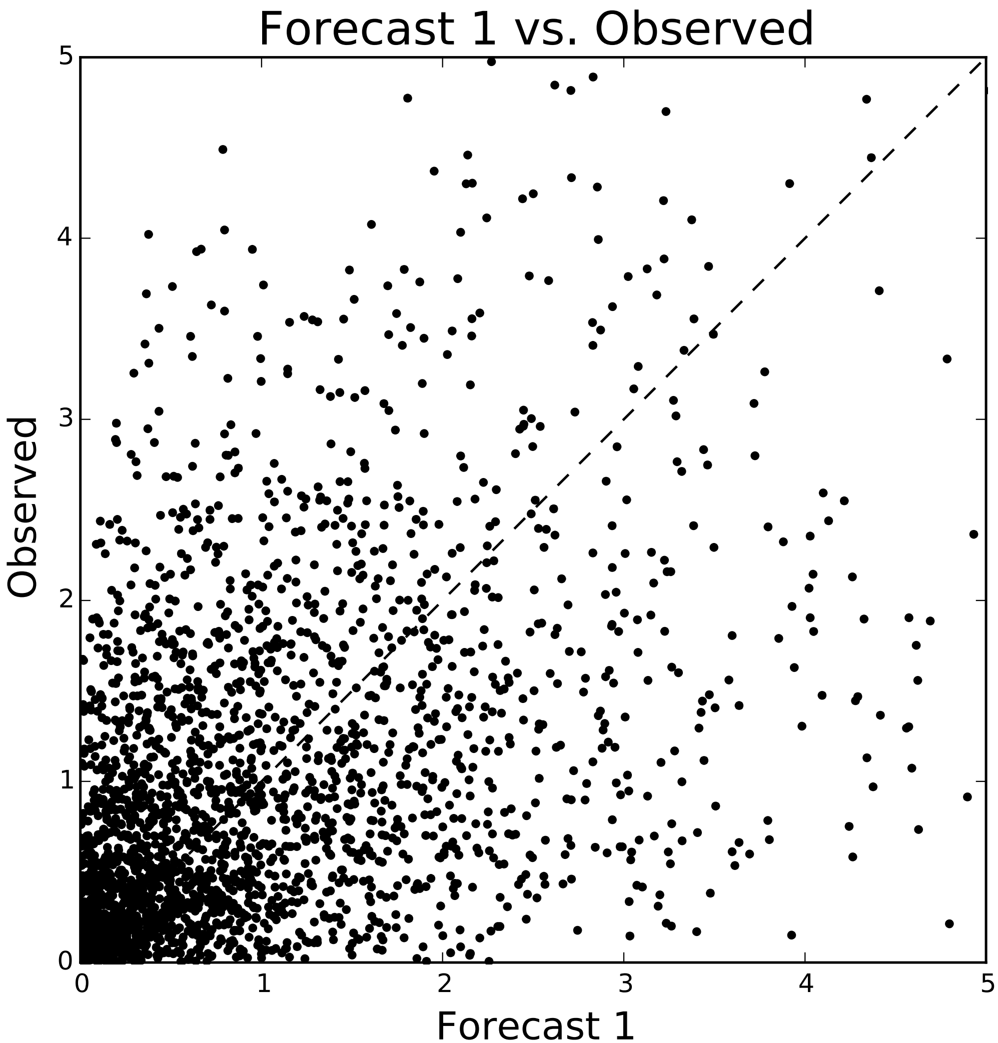
What is the “copula” structure, i.e., the joint probabilities?



a problem when marginal & joint probs. are not well forecast

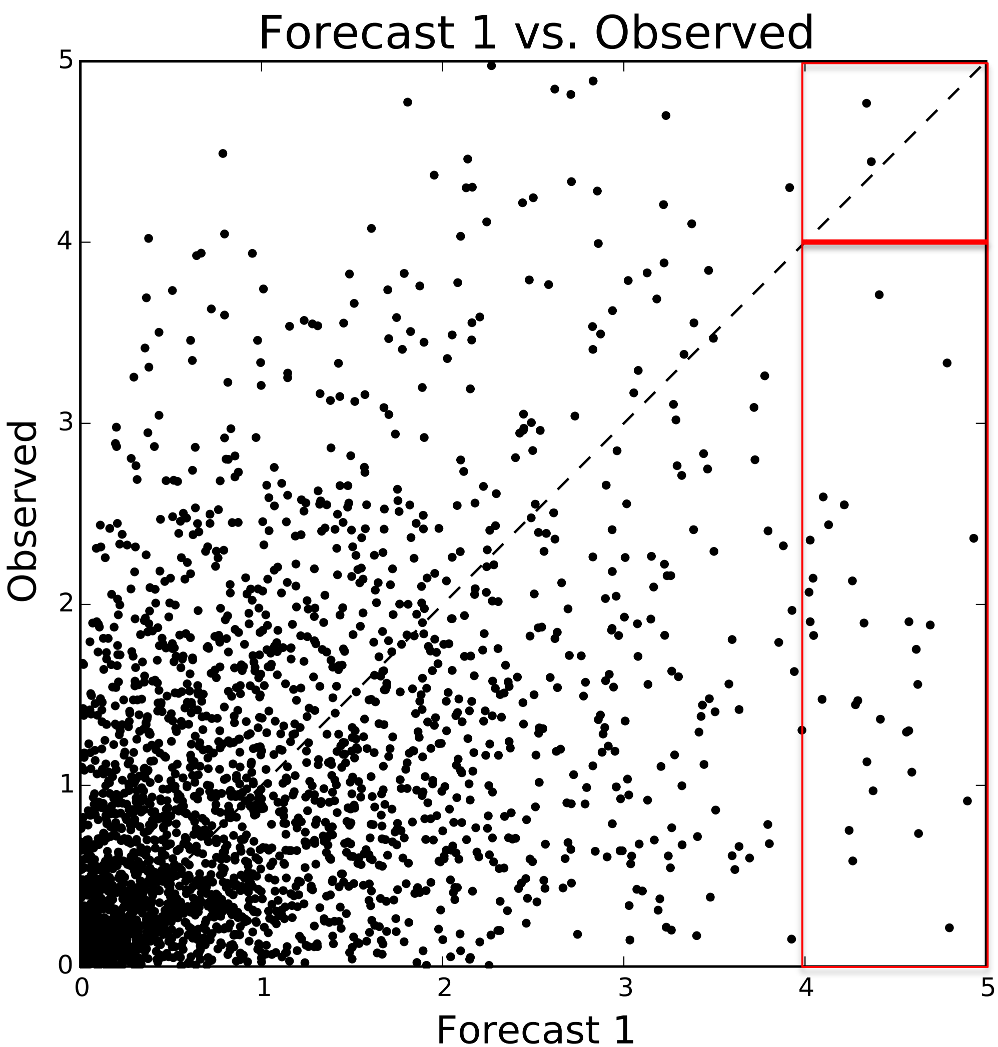
(thanks to conversations with John Schaake)

Simulation study: observations & two forecasts



Suppose we want to know the probability that the obs > 4.0 | forecast 1 > 4.0

Simulation study: observations & two forecasts

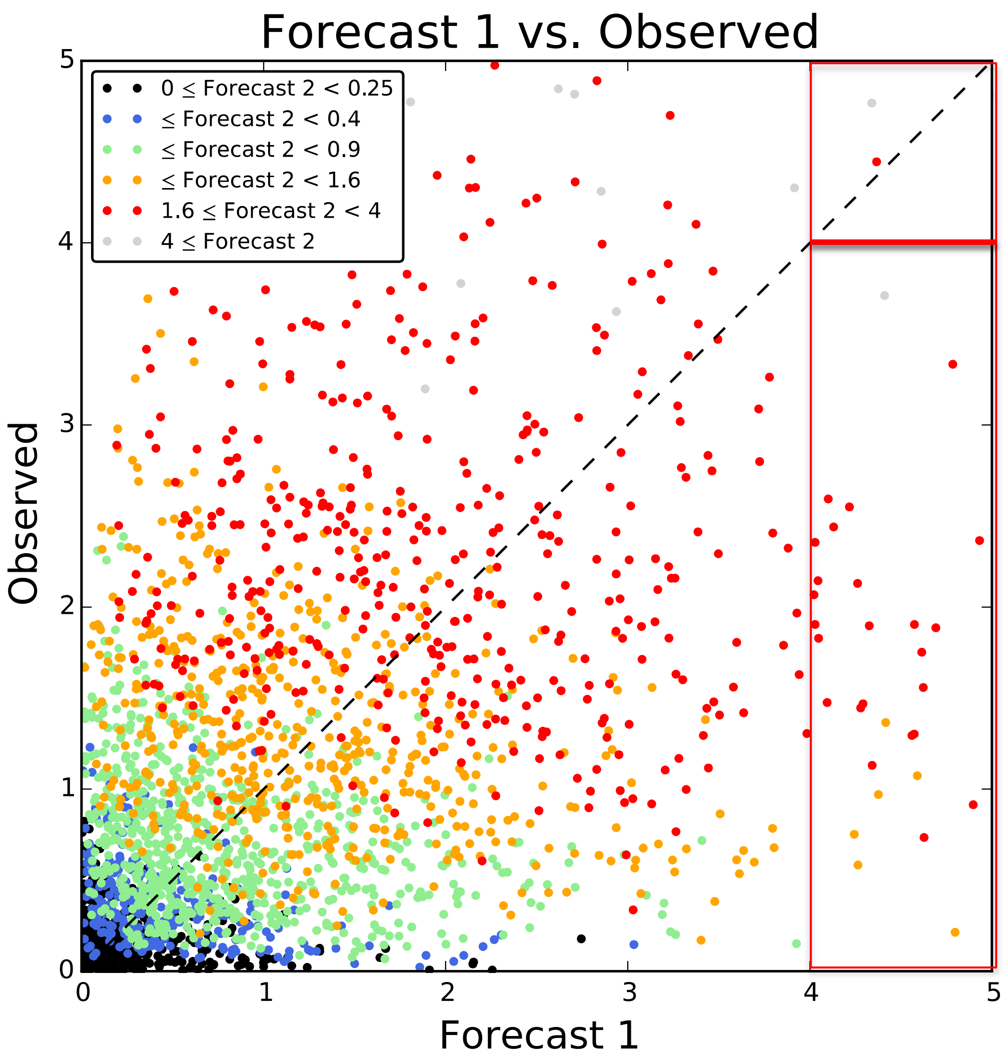


Suppose we want to know the probability that the obs > 4.0 | forecast 1 > 4.0.

We can make some crude estimation from counting:

$$\frac{\#\ \text{fcsts} > 4.0\ \ \text{AND}\ \#\ \text{obs} > 4.0}{\#\ \text{fcsts} > 4.0}$$

Simulation study: observations & two forecasts

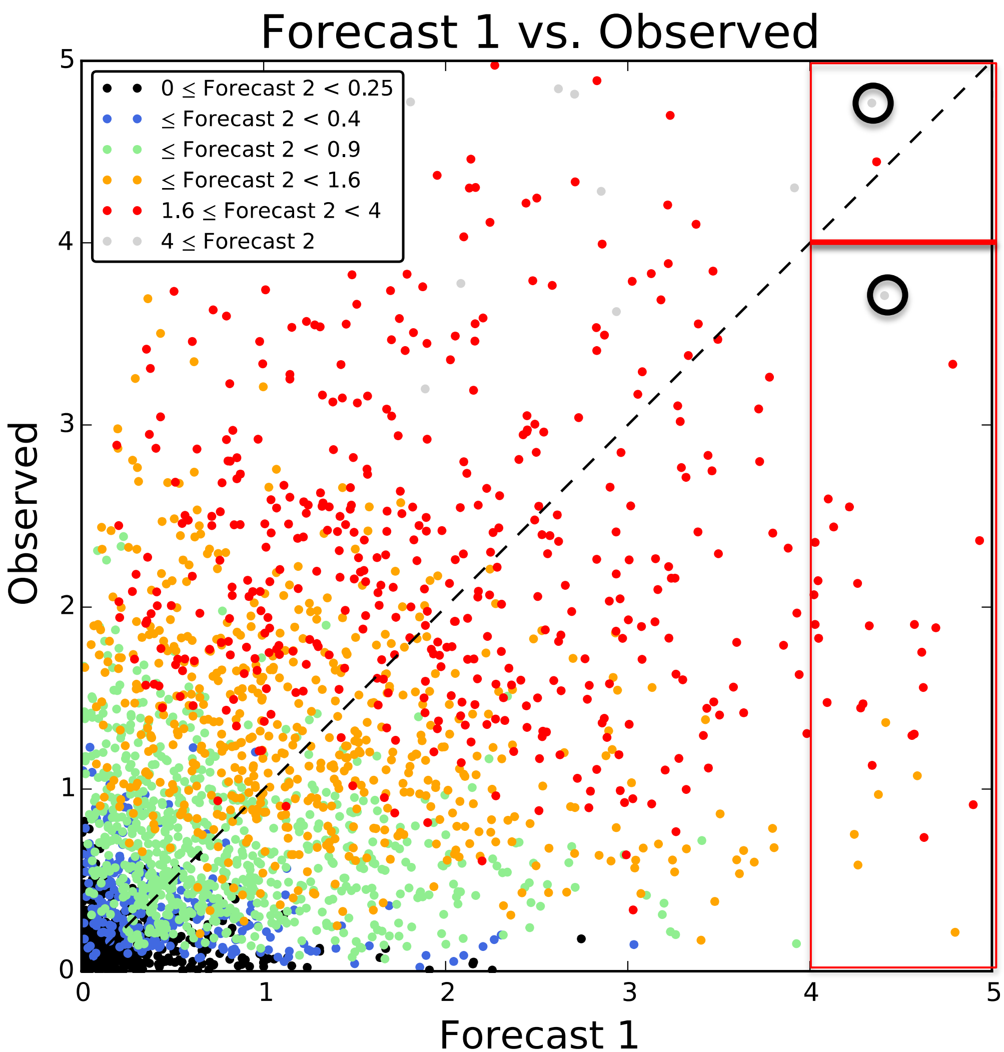


Suppose we want to know the probability that the obs > 4.0 | f1 AND f2 > 4.0.

We can make some estimation from counting (grey dots):

$$\frac{\#\ f1 > 4.0\ \text{ AND } f2 > 4.0 \text{ AND } \#\ obs > 4.0}{\#\ f1 > 4.0\ \text{ AND } f2 > 4.0}$$

Simulation study: observations & two forecasts



Suppose we want to know the probability that the obs > 4.0 | f1 and f2 > 4.0.

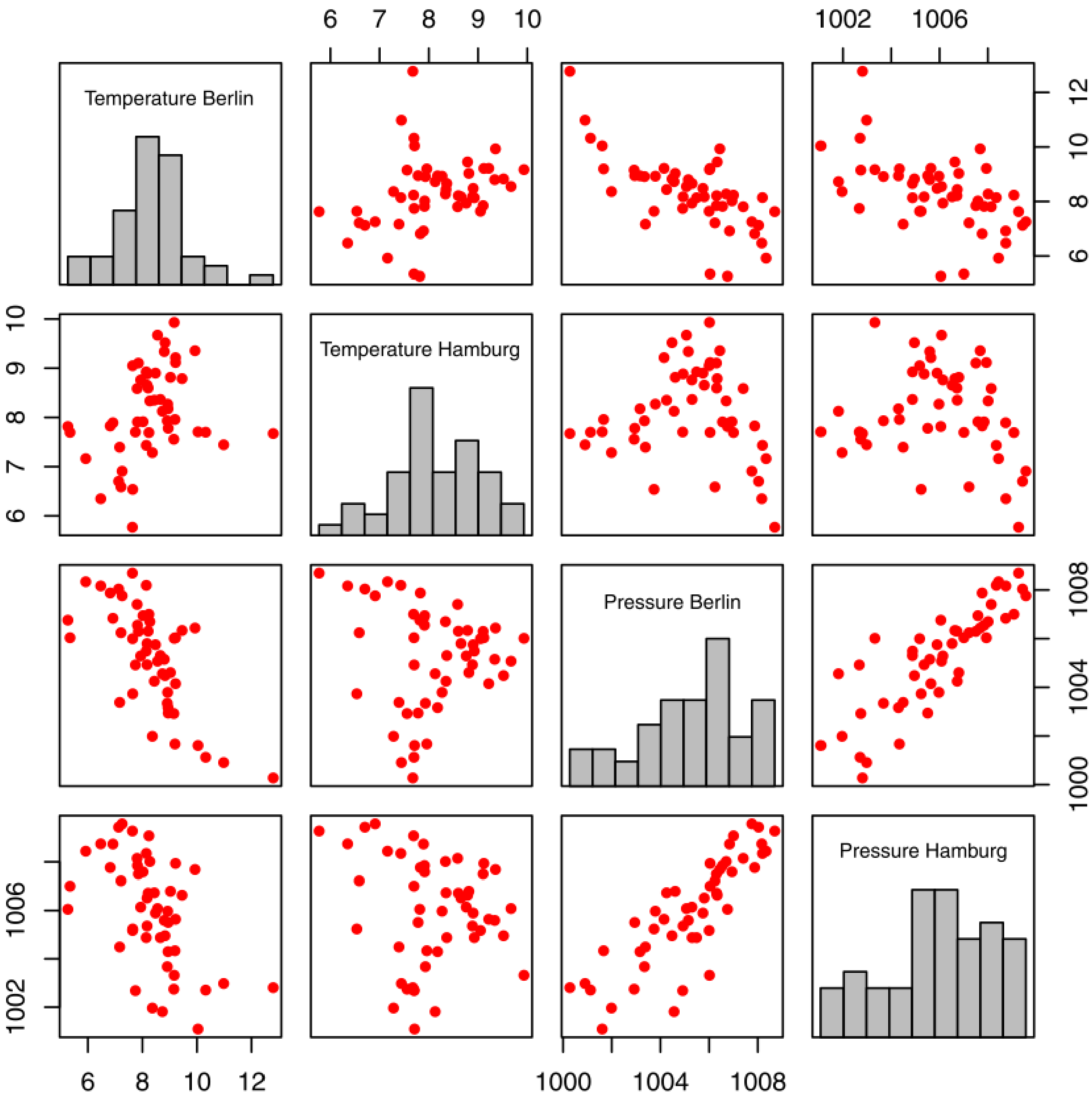
We can make some estimation from counting (grey dots):

$$\frac{\#\,f1 > 4.0 \text{ AND } f2 > 4.0 \text{ AND } \#\,obs > 4.0}{\#\,f1 > 4.0 \text{ AND } f2 > 4.0}$$

This under-sampling problem gets worse and worse with higher and higher dimension; the *"curse of dimensionality."*

Estimating joint probabilities and ameliorating curse of dimensionality.

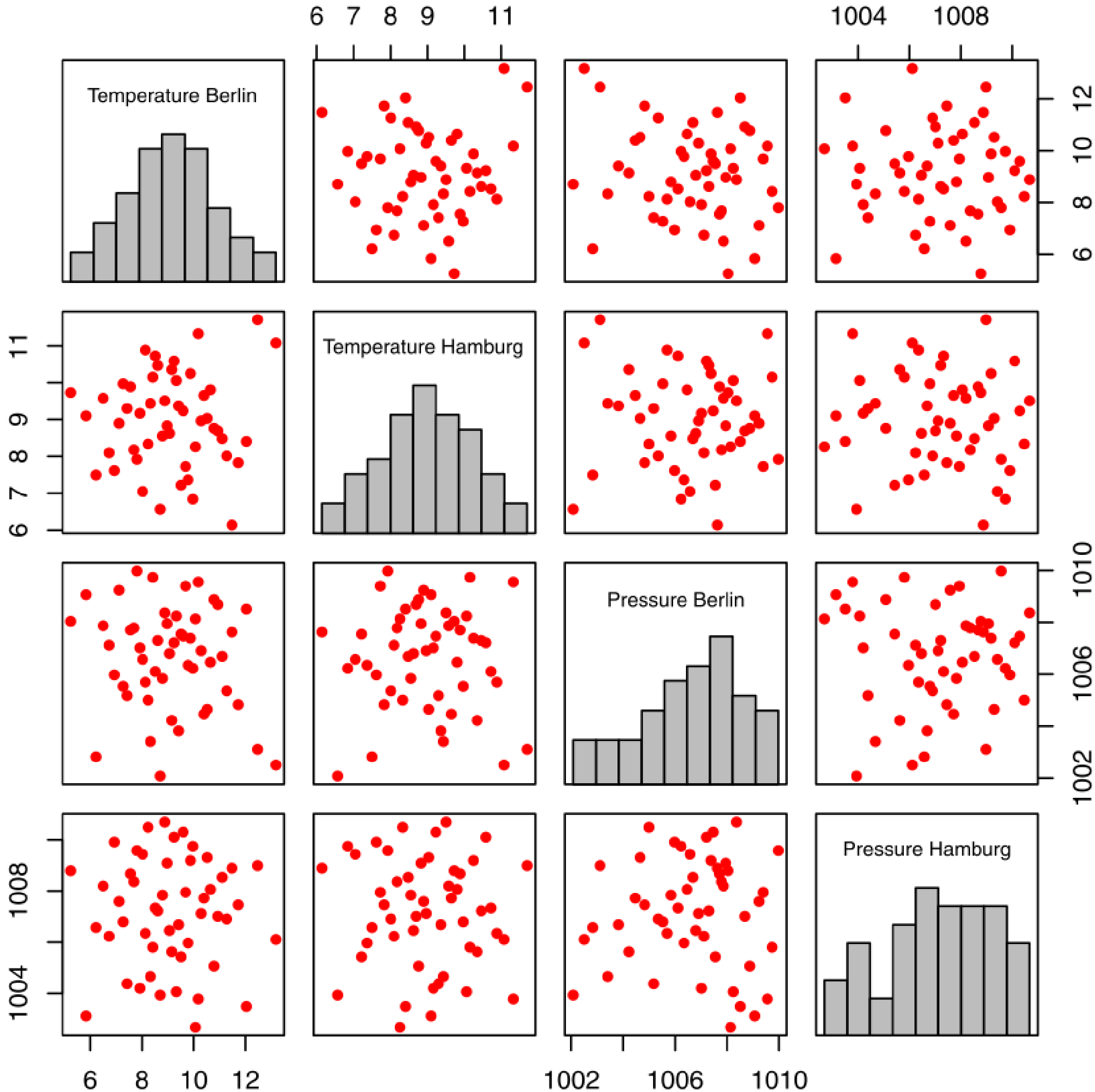
- More training samples.
- Model them parametrically.
  - Suppose the joint probabilities depend on the spatial characteristics of weather forecast, e.g., scattered heavy rain vs. widespread heavy?
  - Then you could sub-divide your training data (scattered batch, widespread batch); must evaluate whether sub-division improves the model more than the reduced sample size degrades it.
- Exploit the joint probability information in the raw ensemble ("*ensemble copula coupling*")?
- Schefzik et al., *Statistical Science* 2013, **28**, 616–640.



(a) Raw ECMWF ensemble

50 raw ECMWF ensemble forecasts of temperature and pressure at two locations.

from Schefzik et al. , Statistical Science 2013, 28, 616–640



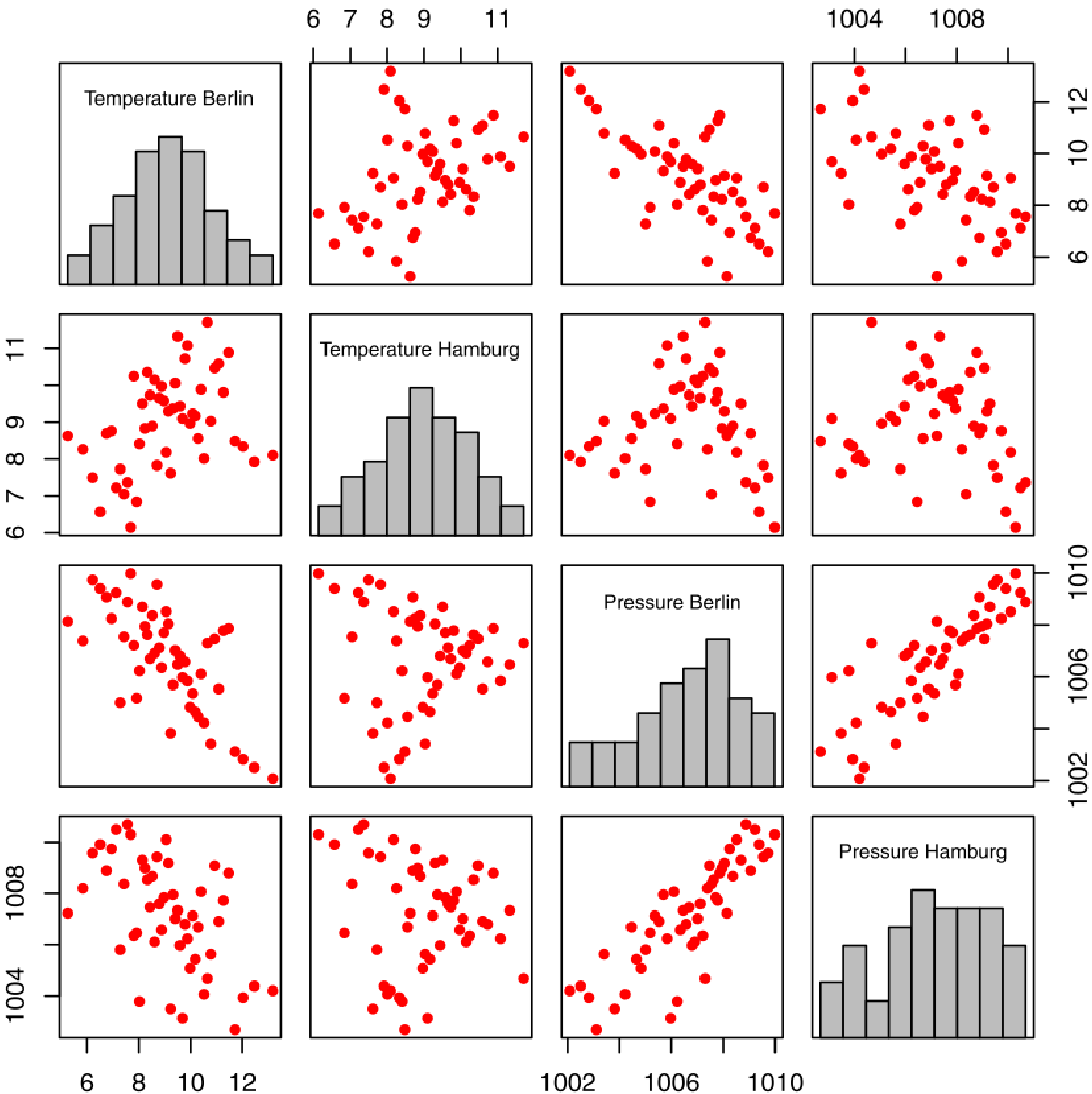
(b) Individual BMA postprocessing

Post-processing takes place independently for each variable using BMA following Fraley et al., MWR, 2010.

50 discrete samples are regenerated independently for each variable from the BMA PDF.

The correlative structure in the raw ensemble is lost.

from Schefzik et al. , Statistical Science 2013, 28, 616–640



(c) ECC postprocessed ensemble

Ensemble copula coupling: rank-order statistics are used to restore the correlative structure of the raw data while preserving the bias and spread corrections produced by BMA.

Q: were those correlations properly estimated by the forecast model?

Wilks (2014; DOI: 10.1002/qj.2414) provides an example of where the "Schaake Shuffle" (climatological covariances) are preferred.

from Schefzik et al. , Statistical Science 2013, 28, 616–640

Building a high-quality statistical model

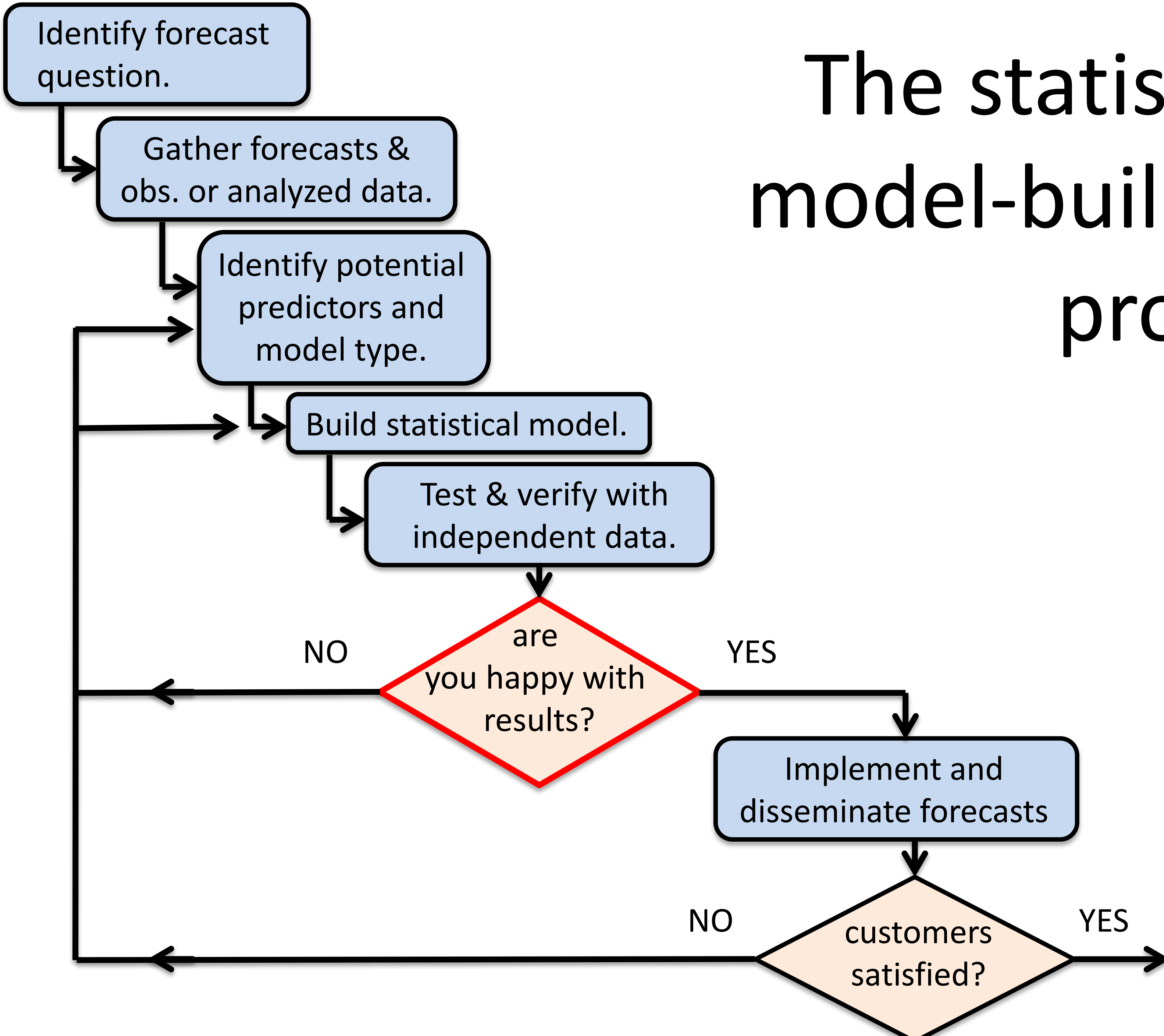
- Old problems are new problems
  - “ *bias-variance tradeoff* ”
  - “ *extrapolating the regression* ”
  - “ *curse of dimensionality* ”
- More modular, reusable software.

Modular software and data library

It's hard to determine whether someone has made an improvement when everyone tests with their own data set, or codes their own version of post-processing methods & verification methods.

Building and supporting a reference library would help our field immensely.

The statistical model-building process



NOAA has lots of test beds.

Aviation Weather Testbed

AWT tests new science and technology to produce better aviation weather products and services.

Climate Testbed

CTB accelerates transition of scientific advances from the climate research community to improved NOAA climate forecast products and services. (Charter)

Coastal & Ocean Modeling Testbed

COMT accelerates transition of advances from the coastal and ocean modeling research community to improved operational ocean products and services. (Charter)

Developmental Testbed Center

DTC improves weather forecasts by facilitating transition of the most promising new NWP techniques from research into operations. (Charter)

GOES-R Proving Gound

GRPG tests and evaluates simulated GOES-R products before the GOES-R satellite is launched into space. (Charter)

Hazardous Weather Testbed

HWT accelerates transition of new meteorological insights and technologies into advances in forecasting and warning for hazardous weather events. (Charter)

Hydrometeorology Testbed

HMT conducts research on precipitation and weather conditions that can lead to flooding, and fosters transition of scientific advances and new tools into forecasting operations. (Charter)

Joint Center for Satellite Data Assimilation

JCSDA accelerates and improves use of research and operational satellite data in weather, ocean, climate and environmental analysis and prediction systems. (Charter)

Joint Hurricane Testbed

JHT is a competitive, peer-reviewed, granting process to choose the best mature research products for testing and transitioning to operations. Includes modeling, data gathering, and decision support components. (Charter)

Operations Proving Ground

OPG serves as a framework to advance NWS decision-support services and science & technology for a weather-ready nation. (Charter)

Space Weather Prediction Testbed

SWPT supports development and transition of new space weather models, products, and services. Infuses new research to improve accuracy, lead-time and value of products, forecasts, alerts, watches, and warnings. (Charter)

Aviation Weather Testbed

AWT tests new science and technology to produce better aviation weather products and services.

Climate Testbed

CTB accelerates transition of scientific advances from the climate research community to improved NOAA climate forecast products and services. (Charter)

Coastal & Ocean Modeling Testbed

COMT accelerates transition of advances from the coastal and ocean modeling research community to improved operational ocean products and services. (Charter)

Developmental Testbed Center

DTC improves weather forecasts by facilitating transition of the most promising new NWP techniques from research into operations. (Charter)

GOES-R Proving Gound

GRPG tests and evaluates simulated GOES-R products before the GOES-R satellite is launched into space. (Charter)

Hazardous Weather Testbed

HWT accelerates transition of new meteorological insights and technologies into advances in forecasting and warning for hazardous weather events. (Charter)

Hydrometeorology Testbed

HMT conducts research on precipitation and weather conditions that can lead to flooding, and fosters transition of scientific advances and new tools into forecasting operations. (Charter)

Joint Center for Satellite Data Assimilation

JCSDA accelerates and improves use of research and operational satellite data in weather, ocean, climate and environmental analysis and prediction systems. (Charter)

Joint Hurricane Testbed

JHT is a competitive, peer-reviewed, granting process to choose the best mature research products for testing and transitioning to operations. Includes modeling, data gathering, and decision support components. (Charter)

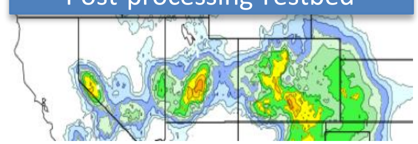
Operations Proving Ground

OPG serves as a framework to advance NWS decision-support services and science & technology for a weather-ready nation. (Charter)

Space Weather Prediction Testbed

SWPT supports development and transition of new space weather models, products, and services. Infuses new research to improve accuracy, lead-time and value of products, forecasts, alerts, watches, and warnings. (Charter)

Post-processing Testbed



PPT supports development of improved decision support tools through the statistical post-processing of numerical weather guidance. It works hand-in-hand with other testbeds.

Is post-processing development handled best when it's dispersed amongst many test beds, by application?

Or is a standalone post-processing test bed with links to other test beds a preferred approach?

Conclusions

- Users increasingly seek more post-processed model guidance; they can't wait for ensembles to become unbiased, perfectly reliable.
- The end product (high quality post-processed guidance) depends on doing each of many steps (data gathering, model selection, evaluation, etc.) well.
- Thorny old statistical challenges still underlie today's impediments to improved forecasts.
- Greater collaboration and sharing will accelerate progress.
- Finally, thanks to Bob Glahn (and many others at MDL) for their pioneering work.

Supplementary slides

Conventional logistic regression

Denoting as p the probability being forecast, a logistic regression takes the form:

$$p = \frac{\exp[f(\mathbf{x})]}{1 + \exp[f(\mathbf{x})]} \quad (1)$$

where $f(\mathbf{x})$ is a linear function of the predictor variables, $\boldsymbol{x}$,

$$f(\mathbf{x}) = b_0 + b_1 x_1 + b_2 x_2 + \cdots + b_K x_K \quad (2)$$

The mathematical form of the logistic regression equation yields 'S-shaped' prediction functions that are strictly bounded on the unit interval ($0 < p < 1$). The name logistic regression follows from the regression equation being linear on the logistic, or log-odds scale:

$$\ln\left[\frac{p}{1-p}\right] = f(\mathbf{x}) \quad (3)$$

from Dan Wilk's article on extended logistic regression, *Met Apps*, 2009

Wilks' extended logistic regression

potentially promising approach is to extend Equations (1) and (3) to include a nondecreasing function $g(q)$ of the threshold quantile q, unifying equations for individual quantiles into a single equation that pertains to any quantile:

$$p(q) = \frac{\exp[f(\mathbf{x}) + g(q)]}{1 + \exp[f(\mathbf{x}) + g(q)]} \qquad (5)$$

or,

$$\ln\left[\frac{p(q)}{1 - p(q)}\right] = f(\mathbf{x}) + g(q) \qquad (6)$$

One interpretation of Equation (6) is that it specifies parallel functions of the predictors $\boldsymbol{x}$, whose intercepts $b_0{}^*(q)$ increase monotonically with the threshold quantile, q:

$$\begin{aligned}\ln\left[\frac{p(q)}{1 - p(q)}\right] &= b_0 + g(q) + b_1 x_1 + b_2 x_2 + \cdots + b_K x_K \\ &= b_0^*(q) + b_1 x_1 + b_2 x_2 + \cdots + b_K x_K \quad (7)\end{aligned}$$

g(q) = SQRT(q); square-root transformation applied to precipitation data as well.