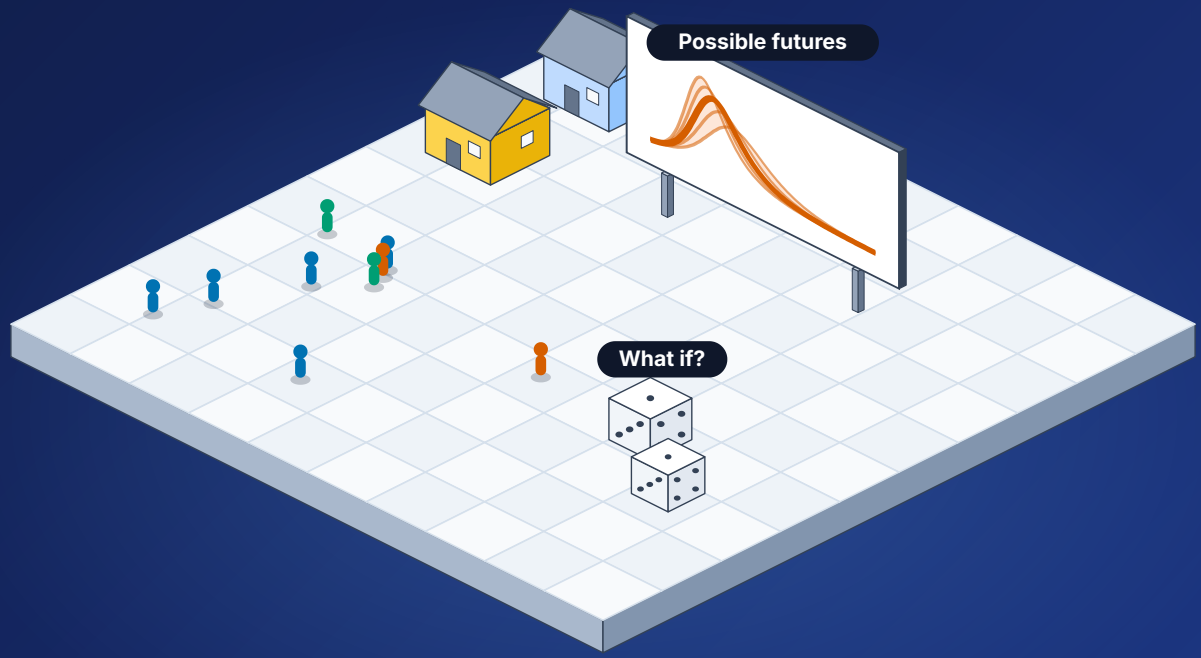
Applied Computational Mathematics Lab

EPIDEMIC MODELING · MODULE 11

Uncertainty, sensitivity and model comparison

Where uncertainty comes from, and how to measure it



4 EPISODES

1. Where uncertainty comes from
2. Identifiability and confounding
3. The sensitivity toolbox: Morris and Sobol
4. Comparing models: scores and coverage

Based on Chapter 11 of T. Lazebnik, Epidemic Modeling: Principles, Algorithms, and Applications (Academic Press, 2026)

acml.teddylazebnik.com

BEFORE WE START

Warm-up: true or false?

CHECK YOURSELF

(a) Run the same epidemic model twice with the same parameters and you get the same epidemic.

(b) A model that fits the first month of cases closely has found the true reporting share.

(c) An input whose first-order Sobol index is near zero can safely be fixed at any value.

(d) Of two forecasters with the same median, the one with narrower intervals always scores better.

HOW TO USE THIS

Guess now; the module settles each one. Answers at the end.

WHY THIS MODULE

A forecast without uncertainty is a guess dressed up as a fact. This module asks where the uncertainty comes from, which inputs drive it, and how to tell a good model from a lucky one.

BEFORE WE START

This module's map

PART 1

Sources of uncertainty

Chance, parameters, observation, model structure, scenarios and computation, added one at a time.

PART 2

Identifiability

When the data cannot tell parameters apart: ties, assumptions, sloppiness, and the data that breaks them.

PART 3

Sensitivity analysis

Local derivatives, Morris screening and Sobol indices: which inputs drive which outputs, and at what cost.

PART 4

Comparing models

Fit against forecast, proper scores, calibration, blends, and fitness for purpose.

GOALS

By the end you should be able to name the sources behind a forecast's spread and propagate them by simulation, spot parameters the data cannot separate and find data that separates them, choose a sensitivity method that fits your question and budget, and compare models with backtests and proper scores.

ONE RULE

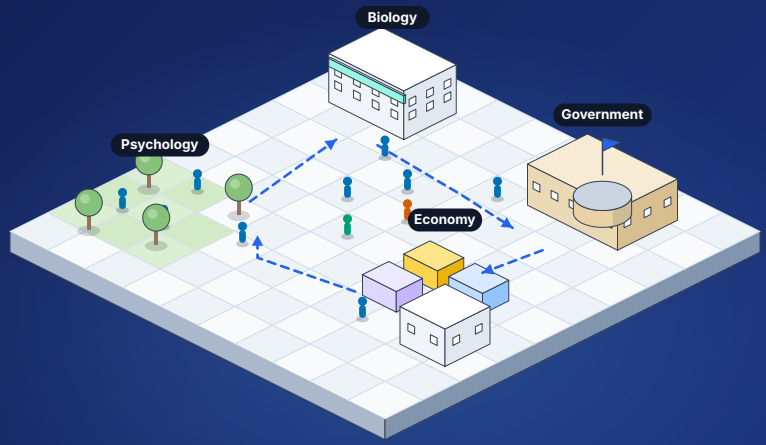
Every number in this deck comes from made-up towns and outbreaks that we simulate, with known truths. The values show mechanisms, not any real disease.

Applied Computational Mathematics Lab

PART 1

Where uncertainty comes from

Six sources, added one at a time.



PART 1 · SIX SOURCES

Where the spread of a forecast comes from

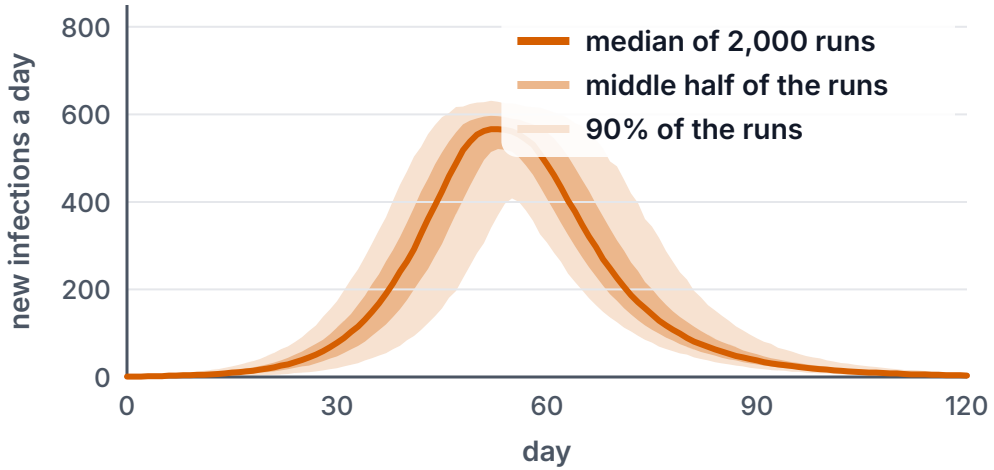
SOURCE	WHAT VARIES	KIND	SHRINKS WITH
chance (process)	who meets whom, who recovers when	aleatory	nothing; it matters less in large populations
parameters	R0, durations, severity	epistemic	better data and studies
observation	reporting share, delays, noise	both	knowing the surveillance system
model structure	compartments, mixing, assumptions	epistemic	comparing models, new knowledge
scenario	future decisions and behavior	choice	nothing; state it as “if …, then …”
computation	samplers, solvers, approximations	numerical	more runs, convergence checks

- **Aleatory**: randomness in the process itself. Better data will not remove it.
- **Epistemic**: lack of knowledge. Better data can shrink it.
- The distinction is practical: it tells you which part of the spread is worth paying to reduce.

IN PLAIN WORDS

An interval is only as honest as the list of sources it includes. Say which ones are in, and which are left out.

Same parameters, different epidemics

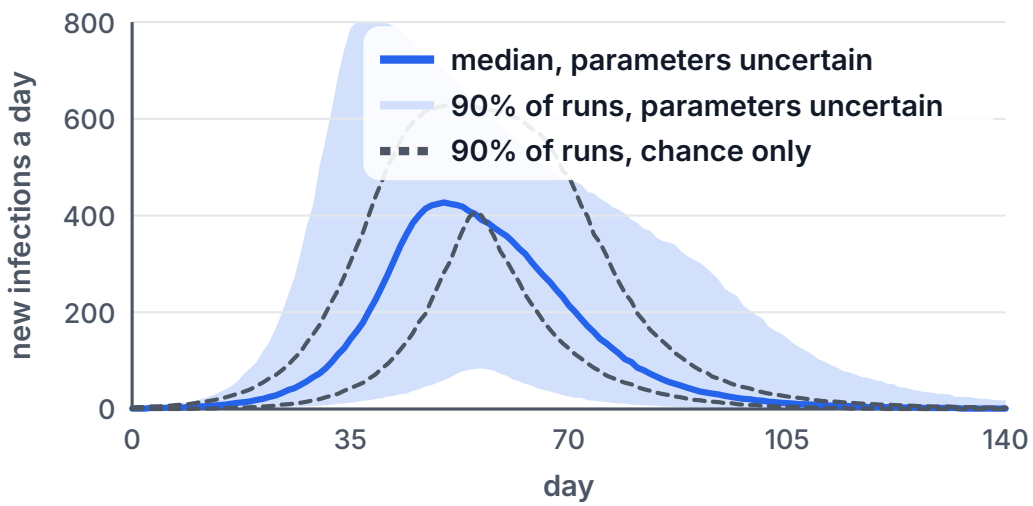


New infections a day in our town of 20,000: a stochastic SIR model (R0 2.5, infectious for 10 days on average, 5 first infections), run 2,000 times.

SAME PARAMETERS, 2,000 RUNS	90% OF RUNS
share ever infected	90.2%–91.3%
day of the peak	47–64
peak, infections a day per 1,000 people	29.7–32.8
runs that fizzled out (under 1% infected)	11

- The final size barely moves: in a town of 20,000 the randomness averages out.
- The timing does move: it is set by the first few chains of infection, while numbers are small.
- This is process uncertainty. It is aleatory: no survey will remove it, and in small populations or at the very start it can dominate.

Not knowing the parameters widens everything



R0 drawn between 2 and 3 and the infectious period between 8 and 12 days (uniform), 2,000 runs; dashed: the 90% band with chance alone.

90% OF RUNS	CHANCE ONLY	+ PARAMETERS
day of the peak	47–64	38–81
peak per 1,000 people	29.7–32.8	20.1–45.1
share ever infected	90.2%–91.3%	83.2%–94.7%

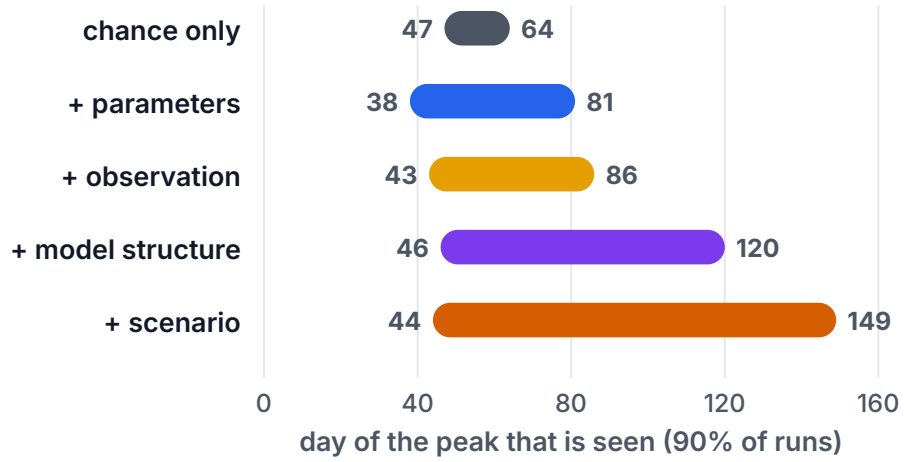
$$\mathrm{Var}(Y) = \mathbb{E}_\theta\big[\mathrm{Var}(Y \mid \theta)\big] + \mathrm{Var}_\theta\big(\mathbb{E}[Y \mid \theta]\big)$$

Peak day, 300 parameter draws with 20 runs each: chance 38 days², parameters 141 days².

- About 79% of the spread in the peak day came from not knowing the parameters, the rest from chance.
- Parameter uncertainty is epistemic: better data and studies can shrink it.

PART 1 · LAYERS

Adding the sources one at a time



Seen: true infections for the first two layers, reported cases from the third layer on. Each layer keeps the ones above it; 2,000 runs each.

LAYER	SEEN PEAK A DAY	SHARE EVER INFECTED
chance only	594–656	90%–91%
+ parameters	402–902	83%–95%
+ observation	134–380	83%–95%
+ model structure	110–342	83%–95%
+ scenario	56–311	56%–94%

- **Observation**: 20–40% of infections reported, 6 days later on average, with extra noise. The reported peak comes about 5 days late and its height spans almost threefold.
- **Structure**: half the analysts add a 3-day latent period with the same R0; the true peak day now spans 41–114.
- **Scenario**: an even chance of a 30% contact cut from day 40; the share ever infected now spans 56%–94%.

PART 1 · PROPAGATION

Propagate it, then report it

MONTE CARLO PROPAGATION

1. Give each uncertain input a distribution: from data, from other studies, or from honest expert ranges.
2. Draw a parameter set; if several model structures are plausible, draw one too.
3. Run the model with its own randomness, then pass the result through the observation model.
4. Repeat thousands of times and report quantiles of the outputs that matter for the decision.
5. Keep scenarios separate (“if this, then that”) unless their probabilities are known.

IN PLAIN WORDS

Leave out a layer and the intervals come out too narrow: the confidence is borrowed.

COMPUTATIONAL UNCERTAINTY

Monte Carlo error, unconverged samplers, solver tolerances and particle filters that collapse onto a few particles all add error, and often fail quietly. Check with repeated runs, convergence diagnostics and tighter tolerances.

WHAT TO REPORT

Which sources are included and which are left out; intervals for the outputs that drive the decision; and which source dominates, so readers know what more data would buy.

Check yourself: Part 1

CHECK YOURSELF

1. Aleatory, epistemic or scenario? (i) How many people one infected wedding guest infects. (ii) The true average infectious period of a new virus. (iii) Whether schools close next month.

2. The total variance of the peak day is 180 days², and the average variance within each parameter set is 40 days². What share of the variance comes from the parameters?

3. A model's intervals for true infections are well calibrated, but you use them for reported cases. Why will they be too narrow, and possibly in the wrong place?

TIP

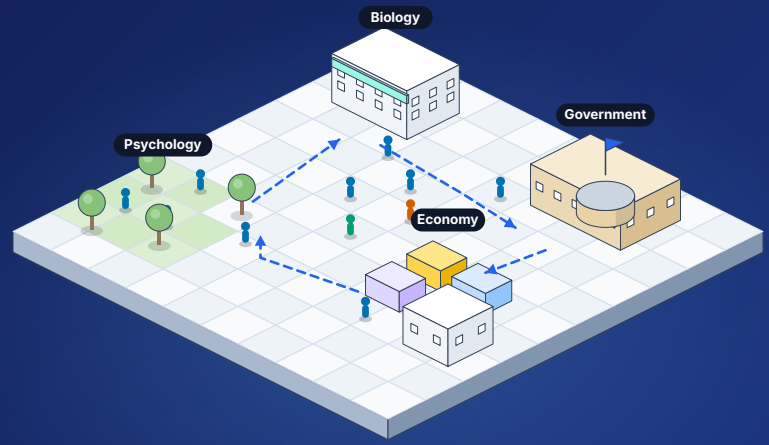
Answers are at the end of the deck. Try them before you look.

Applied
Computational
Mathematics
Lab

PART 2

Identifiability and confounding

When the data cannot tell parameters apart.



PART 2 · TWO KINDS

Can the data pin the parameters down?

STRUCTURAL IDENTIFIABILITY

With perfect, noise-free data from the model, different parameter values always give different outputs.

PRACTICAL IDENTIFIABILITY

With the finite, noisy data we actually have, the fit is peaked enough to give useful ranges.

$$c(t) \approx \rho\, I_0\, e^{rt} \quad \text{(reported cases, early on)}$$

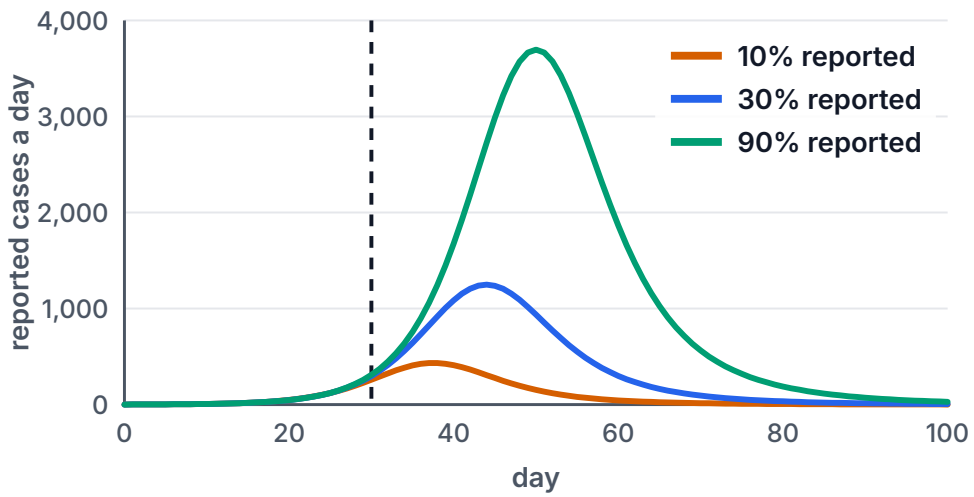
- Early in an outbreak the data fix the growth rate r and the product ρI_0, but not the reporting share ρ or the first infections I_0 separately.
- Two parameters, or a parameter and an assumption, are **confounded** when they can trade off and still give the same data.
- Structural questions can be settled with symbolic tools based on differential algebra; practical ones with profile likelihoods and the Fisher information (Raue et al., 2009).
- A good fit is not evidence that the parameters are known. It only says that some combination of them is.

IN PLAIN WORDS

Identifiability is the bridge between fitting (Module 6) and uncertainty: it tells you which uncertainty the data can remove.

PART 2 · A TIE

Same first month, very different futures



Reported cases in three SIR towns of 100,000 (infectious for 10 days); the transmission rate and the first infections fitted so that days 0–29 (left of the dashed line) match the middle town.

REPORTING SHARE	FIRST INFECTIONS	LARGEST GAP, DAYS 0–29	PEAK DAY	TOTAL REPORTED
10%	27.5	9.3%	38	9,463
30%	10	0.0%	44	28,256
90%	3.4	3.0%	50	84,630

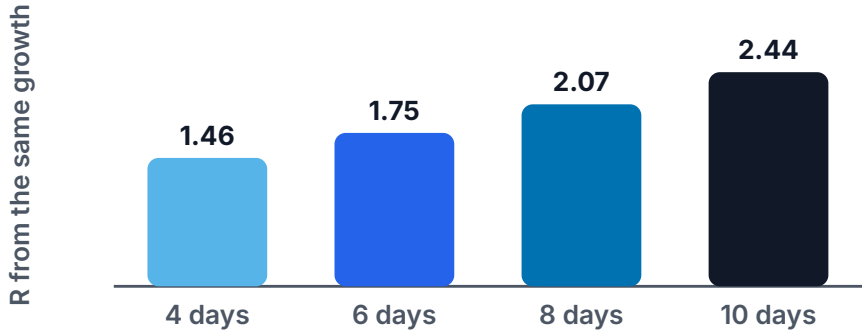
The largest gap is under 10%, smaller than the day-to-day noise in real counts.

- The first 30 days fix the growth and the product of reporting and infections; they cannot separate the towns.
- Their futures differ: reported peaks on days 38, 44 and 50, and a ninefold range in what is still to come.
- Later data help: the peak only arrives when enough people are immune, and that depends on the true number infected.

The generation interval turns growth into R

$$R = \frac{1}{M(-r)}, \qquad \text{gamma interval (shape } k\text{, scale } \theta\text{):} \quad R = (1 + r\theta)^k$$

M is the moment-generating function of the generation interval (Wallinga and Lipsitch, 2007).



Our runs: growth rate 0.1 a day (doubling every 6.9 days); gamma-shaped intervals with a standard deviation of half the mean.

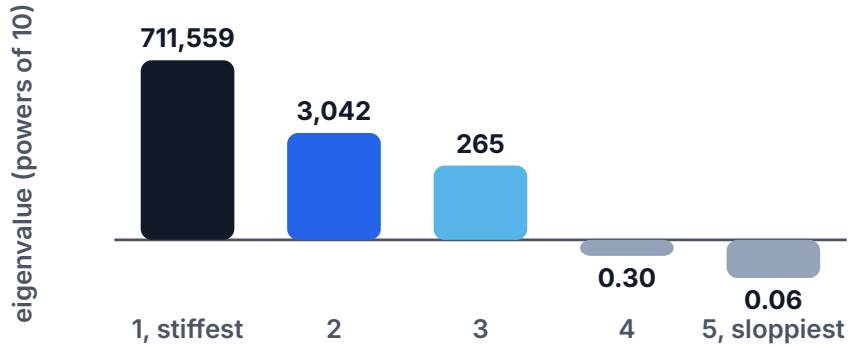
- The same epidemic gives R from 1.46 to 2.44: the assumed interval, not the case counts, makes the difference.
- It matters for policy: the immune share needed to stop growth, $1 - 1/R$, is 32% with a 4-day interval and 59% with a 10-day one.
- The interval itself can be estimated from pairs of linked cases (who infected whom), which is data the growth curve does not contain.

IN PLAIN WORDS

Assumptions confound too. State them, and show how the answer moves when they change.

Identifiable, but barely: sloppy models

$$\mathcal{I}(\theta) = \sum_t \frac{1}{\mu_t} \frac{\partial \mu_t}{\partial \theta} \left(\frac{\partial \mu_t}{\partial \theta} \right)^\top, \qquad \mathrm{sd}(v^\top \theta) \approx \frac{1}{\sqrt{\lambda}}$$



Eigenvalues of the Fisher information: an SEIR model with five parameters in logs, fitted to 60 days of Poisson counts (our runs).

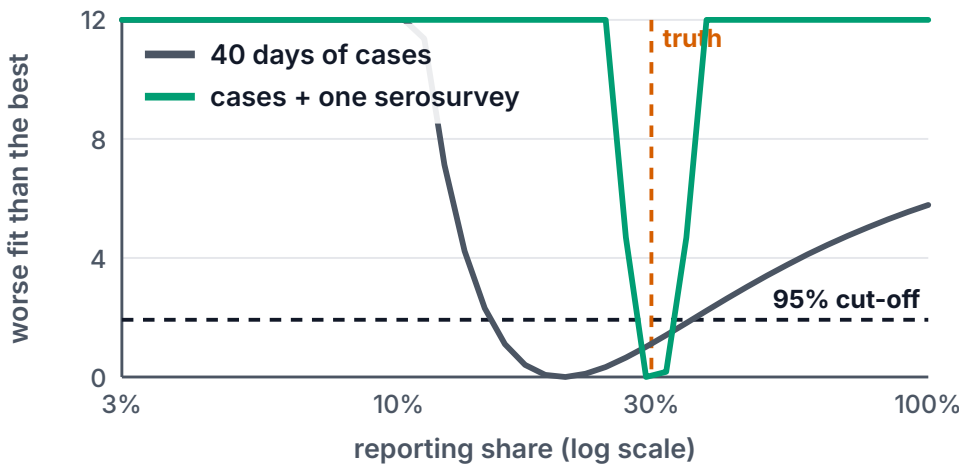
PARAMETER (LOG)	STIFFEST DIRECTION	SD IF FITTED ALONE
transmission rate	-0.84	1.98
rate of leaving latency	-0.24	1.71
recovery rate	+0.47	3.27
reporting share	-0.13	0.60
first infections	-0.08	1.72

sd: Cramér–Rao standard deviation of each log parameter, with all the others free.

- The largest eigenvalue is 1.25×10^7 times the smallest: the loosest combination is about 3,500 times more uncertain than the stiffest.
- The stiff direction is close to transmission minus recovery: the growth rate, which the data see clearly.
- This is sloppiness, common in many-parameter models (Gutenkunst et al., 2007): predictions that depend only on stiff combinations can still be precise.

PART 2 · BREAKING THE TIE

Data that sees what cases cannot



Profile of the reporting share: for each value, the transmission rate and first infections re-fitted to 40 days of noisy cases (negative binomial); our runs.

PROFILE LIKELIHOOD

1. Fix the parameter of interest at one value on a grid.
2. Re-fit all the others; record the best negative log-likelihood.
3. Keep the values within 1.92 of the best one (95%, one parameter).
4. A flat profile means the parameter is practically unidentifiable.

DATA	95% RANGE, REPORTING SHARE
40 days of cases	15.9%–34.9%
+ a serosurvey of 2,000 on day 40 (506 positive)	29.3%–32.0%
the truth	30%

IN PLAIN WORDS

The serosurvey counts infections directly, so it separates the reporting share from the number infected: the range shrank about sevenfold.

PART 2 · REMEDIES

What to do when parameters are confounded

ADD DATA

Another stream

Serosurveys, hospital records or contact studies see quantities that case counts cannot.

BORROW KNOWLEDGE

Honest priors

Values from other studies, with their uncertainty, rather than a single fixed number.

REPARAMETERIZE

Estimate what is seen

Fit identifiable combinations, such as the growth rate or the product of reporting and infections.

REPORT RANGES

Show the trade-off

Give profile or posterior ranges, and how each prediction moves along the unknown direction.

DESIGN

Choose the next data

Collect the data that would shrink the uncertainty of the prediction the decision depends on.

IN PLAIN WORDS

Before you trust a parameter, ask: could different values tell the same story? If so, find data that tells them apart, or report the whole range.

A QUICK CHECK

Two parameters fitted to one curve will often trade off. Plotting the fit's best values of one against the other, over many re-fits to noisy copies of the data, shows the ridge directly.

Check yourself: Part 2

CHECK YOURSELF

4. On day 0 there were 20 reported cases, doubling every 5 days. Can you tell how many infections there were on day 0? What would you need?

5. With a growth rate of 0.1 a day and an exponential generation interval ($k = 1$) of mean 5 days, $R = 1 + r\theta = 1.5$. What is R with a mean of 8 days?

6. A Fisher information has eigenvalues 10^6 and 1. How much more uncertain is the loosest combination than the stiffest?

TIP

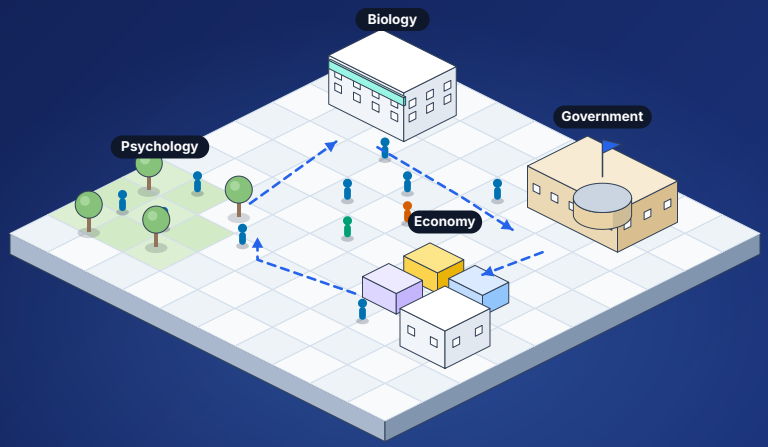
Answers are at the end of the deck.

Applied
Computational
Mathematics
Lab

PART 3

Sensitivity analysis

Which inputs drive which outputs, and at what cost.



PART 3 · THE SET-UP

Seven inputs, three outputs

INPUT	RANGE	BASELINE
R0	1.5 to 3.5	2.5
latent period	2 to 5	3
infectious period	3 to 8	5
reporting share	0.1 to 0.5	0.3
first infections	1 to 50	10
cut in contacts	0 to 0.6	0.3
day the cut starts	20 to 80	50

Our town: SEIR, 100,000 people, one 60-day contact cut. Made-up ranges, drawn uniformly. Shares as fractions.

OUTPUT	AT THE BASELINE	90% OVER THE RANGES
share ever infected	75.8%	60%–96%
peak of reported cases a day	678	157–1,930
day of that peak	49	33–184

- **Factor prioritization**: which input, if known exactly, would shrink the output's uncertainty the most?
- **Factor fixing**: which inputs can be set anywhere in their range without changing the answer?
- **Interactions**: which inputs matter only in combination?

IN PLAIN WORDS

The answer depends on the output you ask about. Choose it from the decision, before running anything.

PART 3 · LOCAL

Nudge one input by 1%

$$\varepsilon_i = \frac{\partial y}{\partial x_i}\,\frac{x_i}{y} \approx \frac{y(1.01\,x_i) - y(0.99\,x_i)}{0.02\,y}$$

INPUT	SHARE INFECTED	REPORTED PEAK	PEAK DAY
R0	+0.90	+6.83	0
latent period	-0.04	-2.01	0
infectious period	-0.06	-3.01	0
reporting share	0	+1.00	0
first infections	+0.02	+0.57	0
cut size	-0.25	0	0
cut start	+0.17	+4.11	+1.02

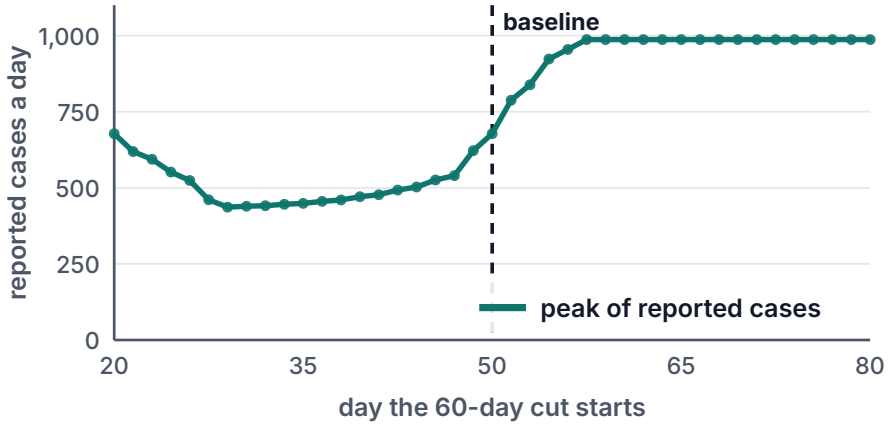
Elasticities at the baseline: % change in the output per 1% change in the input; 15 runs.

- R0: +1% raises the share ever infected by 0.9% and the reported peak by 6.8%.
- The reporting share scales what we see: elasticity 1 for the reported peak, 0 for the share infected.
- Longer latent and infectious periods slow the epidemic and lower its peak.
- The peak day barely moves for a 1% nudge: whole days are coarse. Only the cut's start moved it.

IN PLAIN WORDS

Cheap and clear: one run per nudge. But it describes only the neighborhood of the point you chose.

A slope at one point, not the whole picture



Peak of reported cases by the day the 60-day cut starts, with every other input at its baseline (our runs).

- At the baseline, day 50, the cut lands right at the peak, so a later start lets the peak through: elasticity 4.1.
- A start near day 29 gives the lowest peak (436 a day); from about day 60 the cut comes after the peak and changes nothing (987).
- Elasticities assume small, separate changes. They miss thresholds, curves and interactions, and they change with the baseline.

TRY IT

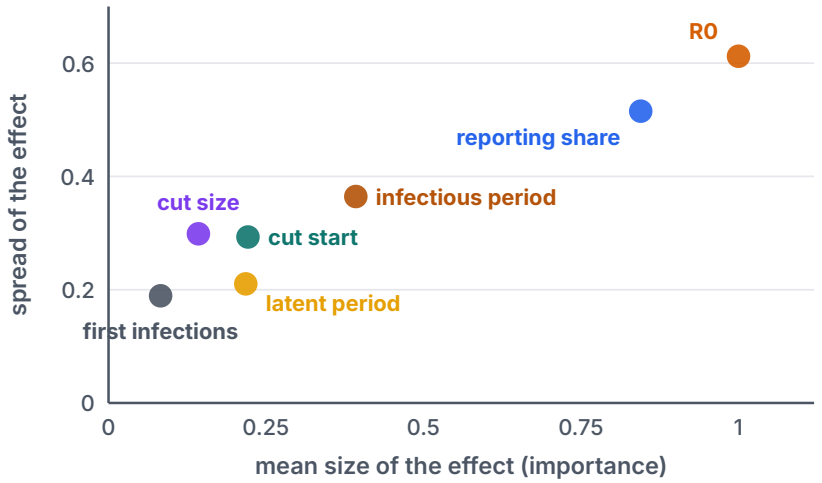
Move the baseline start to day 70. What happens to the local effect of the start day on the reported peak? Is the start day now unimportant?

Morris screening: cheap and global

ELEMENTARY EFFECTS (MORRIS, 1991)

1. Put each input on a grid of p levels across its range (here 8).
2. From a random start, change one input at a time by a fixed step until all have moved: k + 1 runs per path.
3. Each change gives an elementary effect $EE_i = \left[y(x + \Delta e_i) - y(x)\right]/\Delta$.
4. Over r paths: μ_i^* = mean of $|EE_i|$ (importance), σ_i = their spread (curvature, interactions).

Our run: 100 paths × 8 runs = 800 runs.

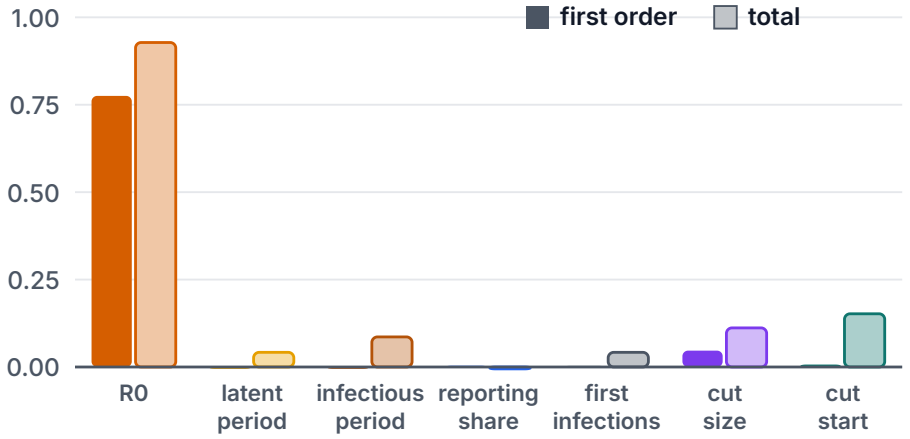


Peak of reported cases; both axes scaled by the largest mean effect.

- Share ever infected: R0 first by far (mean effect 0.36, next 0.12); the reporting share has none.
- Reported peak: R0 first, the reporting share a close second.
- A spread as large as the mean flags curves or interactions, but does not say which.

Sobol indices split the variance

$$S_i = \frac{\mathrm{Var}\big(\mathbb{E}[Y \mid X_i]\big)}{\mathrm{Var}(Y)}, \qquad S_{T_i} = \frac{\mathbb{E}\big[\mathrm{Var}(Y \mid X_{\sim i})\big]}{\mathrm{Var}(Y)}$$



Share ever infected: first-order and total indices; N = 16,384 base samples, 147,456 runs (our runs).

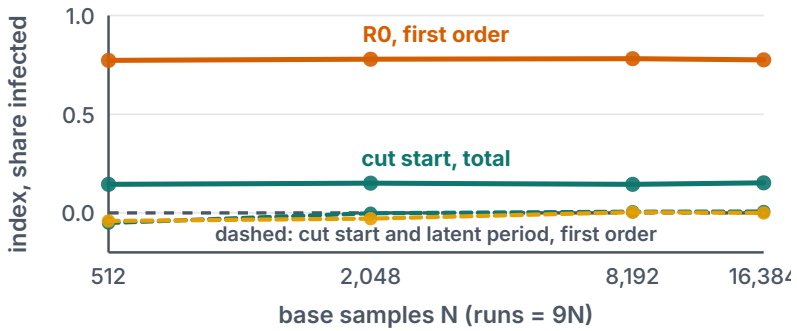
- First order S_i: the share of the variance due to the input alone. Total S_{T_i}: adds every interaction it takes part in.
- $S_{T_i} - S_i$ measures interactions; $\sum_i S_i = 1$ only for additive models.
- R0 alone explains 78% of the variance in the share ever infected. The cut's start explains 0.7% alone, but 15% with its interactions.
- Estimation: two random input matrices A and B, plus k copies of A with one column taken from B, so N(k + 2) runs (Saltelli et al., 2010; Jansen, 1999).

PART 3 · ALL OUTPUTS

Sensitivity depends on the question

INDEX	R0	LATENT PERIOD	INFECTIOUS PERIOD	REPORTING SHARE	FIRST INFECTIONS	CUT SIZE	CUT START
share infected, first order	0.78	0.00	0.00	0.00	0.00	0.05	0.01
share infected, total	0.93	0.04	0.09	0.00	0.04	0.11	0.15
reported peak, first order	0.38	0.03	0.06	0.27	0.00	0.01	0.02
reported peak, total	0.56	0.07	0.14	0.37	0.03	0.04	0.12
peak day, first order	0.60	0.04	0.09	0.00	0.04	0.04	0.03
peak day, total	0.76	0.09	0.17	0.00	0.11	0.14	0.14

N = 16,384; bootstrap 90% ranges lie within ±0.015 of each value; tiny negative estimates shown as 0.



Estimates as the sample grows: small samples give noisy, even negative, first-order values.

- The reporting share: total index 0 for the share infected and the peak day, 37% for the reported peak.
- R0 dominates every output, but its share ranges from three quarters of the variance to under half.
- An input with a tiny first-order index can still matter through interactions: never fix an input on its first-order index alone.

PART 3 · TOOLBOX

Pick the method by budget and question

METHOD	RUNS HERE	TELLS YOU	MISSES
local derivatives	15	slopes at one point	thresholds, interactions, other baselines
one-at-a-time sweeps	287	the shape along each axis	interactions
Morris screening	800	a ranking, and a flag for curvature	how much variance each input explains
Sobol indices	147,456	variance shares, alone and in total	nothing major, but needs a fast model
emulator, then Sobol	a few hundred to train	Sobol indices of a stand-in model	the emulator's own error

An emulator (for example a Gaussian process or a polynomial chaos expansion) is fitted to a few hundred runs and then queried instead of the model.

- A first look: local derivatives. Many inputs: screen with Morris and drop the unimportant ones. The few left: Sobol.
- Slow agent-based models: train an emulator, check it on held-out runs, then analyze the emulator.
- Report the ranges used: sensitivity results are conditional on them.

IN PLAIN WORDS

Sensitivity tells you where to spend effort: which input to measure better, and which to stop arguing about.

Check yourself: Part 3

CHECK YOURSELF

7. An input has a first-order index of 0.01 and a total index of 0.20. What does that tell you? Can you fix it at its baseline?

8. With $k = 10$ inputs and $N = 4096$ base samples, how many model runs does the Saltelli design need?

9. Why can the local elasticity of the cut's start day give a misleading picture of its importance?

TIP

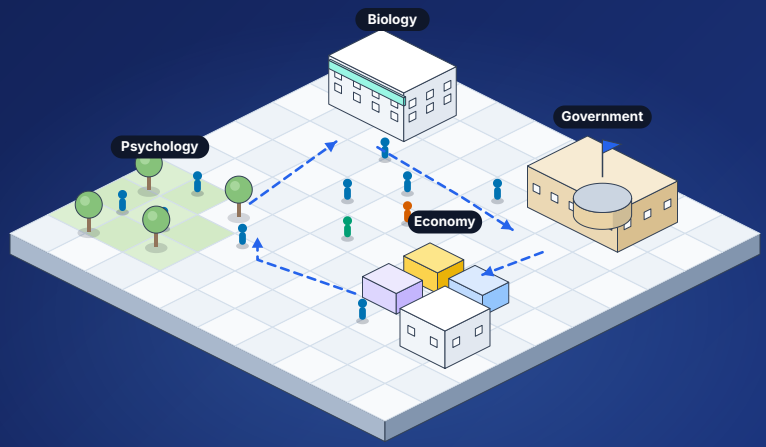
Answers are at the end of the deck.

Applied
Computational
Mathematics
Lab

PART 4

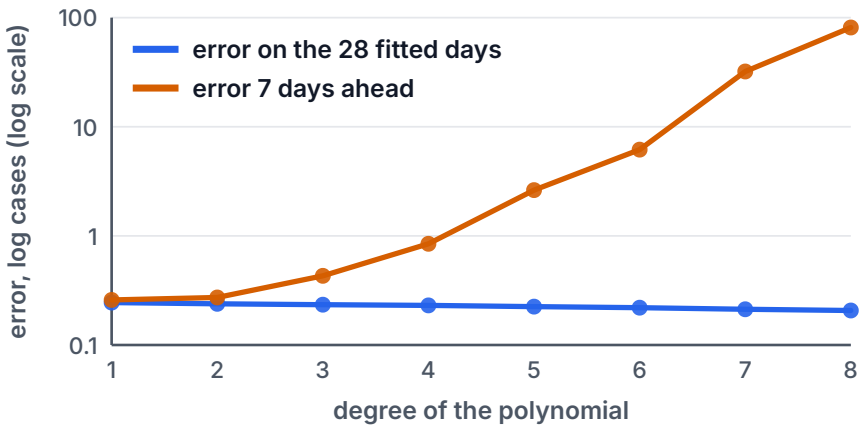
Comparing models

Best for what?



PART 4 · FIT IS NOT FORECAST

A better fit to the past, worse forecasts



Polynomials fitted to 28 days of log cases from a simulated outbreak, from 22 dates; median errors (our runs).

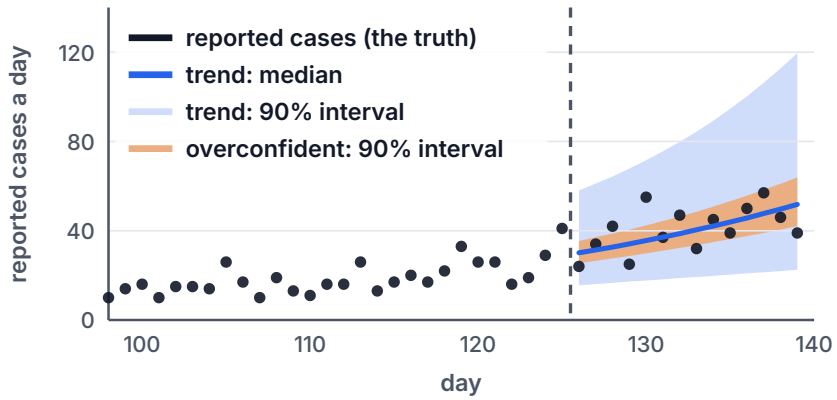
DEGREE	ERROR, FITTED DAYS	AIC	ERROR 7 DAYS AHEAD
1	0.245	-75.0	0.26
2	0.238	-74.4	0.27
3	0.234	-73.5	0.43
4	0.230	-72.4	0.84
5	0.224	-72.0	2.63
6	0.219	-71.2	6.16
7	0.213	-71.0	32.14
8	0.207	-70.5	81.14

Errors in log cases; an error of 0.26 means off by a factor of about 1.3.

- Every extra degree fitted the past better and forecast worse; by degree 5 forecasts were off by a factor of about 14.
- $\mathrm{AIC} = 2p - 2\ln\hat{L}$ penalizes the number of parameters p; it chose the straight line. Better still: test models the way they will be used.

PART 4 · BACKTESTING

Forecast the past using only what was known then



Truth: a simulated town of 500,000 with two waves, 30% of infections reported with noise. Forecasts made on day 126; both have the same median.

ROLLING-ORIGIN BACKTEST

1. Pick origin dates: here 26, one a week from day 42.
2. On each, fit using only the data available before that date: no later revisions, no future.
3. Forecast 1, 7 and 14 days ahead as a set of quantiles (here 21).
4. Score against what happened, then average over origins.

FORECASTER	HOW IT WORKS
last week	last week's mean, flat
trend	a straight line through 14 days of log cases
overconfident trend	the same, intervals 4 times narrower
very wide, flat	last week's mean, very uncertain
blend	the average of the first two

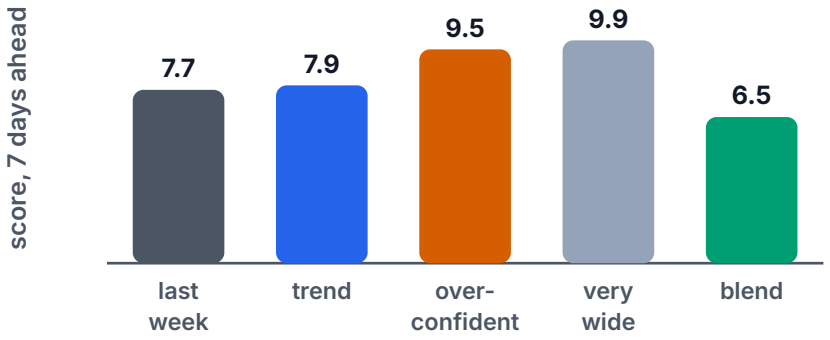
PART 4 · SCORES

Score the whole forecast, not just its middle

$$\mathrm{IS}_\alpha(l,u;y) = (u-l) + \tfrac{2}{\alpha}(l-y)\,1\{y<l\} + \tfrac{2}{\alpha}(y-u)\,1\{y>u\}$$

$$\mathrm{WIS} = \frac{1}{K+\frac{1}{2}}\left(\tfrac{1}{2}\,|y-m| + \sum_{k=1}^{K}\tfrac{\alpha_k}{2}\,\mathrm{IS}_{\alpha_k}\right)$$

Weighted interval score over K central intervals and the median m; lower is better (Bracher et al., 2021).



7 DAYS AHEAD	ABS. ERROR	SCORE	50% COV.	90% COV.	90% WIDTH
last week	10.9	7.7	69%	88%	48
trend	11.9	7.9	62%	96%	63
overconfident trend	11.9	9.5	12%	23%	15
very wide, flat	12.5	9.9	81%	100%	96
blend of the first two	9.0	6.5	65%	96%	55

- The trend and its overconfident copy share a median, so their absolute errors match (11.9). But the copy's 90% intervals caught the truth 23% of the time.
- The score punished that: 9.5 against 7.9. The very wide forecaster caught everything and still scored 9.9.
- A proper score is minimized, on average, by forecasting what you believe; it cannot be gamed by hedging (Gneiting and Raftery, 2007).

Calibration, horizons and blends

SCORE (90% COVERAGE)	1 DAY	7 DAYS	14 DAYS
last week	5.8 (96%)	7.7 (88%)	10.6 (85%)
trend	6.4 (92%)	7.9 (96%)	15.1 (85%)
overconfident trend	7.3 (35%)	9.5 (23%)	19.2 (27%)
very wide, flat	8.8 (100%)	9.9 (100%)	11.4 (100%)
blend of the first two	5.9 (96%)	6.5 (96%)	10.3 (88%)

Our backtest: 26 origins for each horizon; lower scores are better.

- The blend scored best at 7 and 14 days (6.5, 10.3); at 1 day it tied with last week.
- The trend held up for a week, then fell apart at 14 days (15.1): every forecaster has a horizon.

CALIBRATION

Intervals mean what they say: 90% intervals contain the truth about 90% of the time. Check coverage at several levels, and a histogram of where each truth falls within its forecast: flat when calibrated, U-shaped when overconfident, a hump when too wide.

IN PRACTICE

In the US COVID-19 Forecast Hub, an ensemble of many teams' forecasts was among the most consistently accurate (Cramer et al., 2022).

PART 4 · FIT FOR PURPOSE

Best for what?

NOWCASTING

The present

Correct for reporting delays; judge by how close early estimates come to the final counts.

FORECASTING

The next weeks

Backtests, proper scores and calibration at the horizons that matter.

SCENARIOS

What if

Plausible mechanisms and assumptions; they cannot be scored as forecasts, because the scenario rarely happens exactly.

EXPLANATION

Why

Parsimony, identifiable parameters, and mechanisms that agree with independent evidence.

DECISIONS

What to do

Costs are uneven: running out of beds costs far more than spare beds, so the upper tail may matter more than the average error.

IN PLAIN WORDS

Uncertainty quantified, sensitivity explained, models compared for their purpose: that is what makes a model trustworthy.

EXAMPLE

A model with a larger average error but well-calibrated upper intervals can be the better tool for planning hospital surge capacity.

Check yourself: Part 4

CHECK YOURSELF

10. A 90% interval is [20, 60] and the truth is 70. Compute the interval score ($\alpha = 0.1$).

11. Forecasters A and B have the same median. A's 90% intervals cover 23% of the truths, B's 96%. Which has the smaller absolute error? Which would a proper score usually prefer?

12. Give a decision for which a model with a larger average error may be the better choice.

TIP

Answers are on the next slides.

ANSWERS

Answers: warm-up and Parts 1–2

- **(a)** False: with the same parameters, the peak day ranged over days 47–64. **(b)** False: towns reporting 10%, 30% and 90% of infections matched the first 30 days within 10%. **(c)** False: the cut's start had a first-order index of 0.007 but a total of 0.15. **(d)** False: the overconfident copy scored 9.5 against 7.9.
- **1.** (i) aleatory (chance); (ii) epistemic (a parameter); (iii) scenario (a choice).
- **2.** $(180 - 40)/180 \approx 0.78$: about 78%.
- **3.** Reported cases add observation uncertainty (unknown reporting share, delays, noise), so they need wider intervals; a reporting share below one and delays also shift the level and timing.
- **4.** No: early data fix only the growth and the product of reporting share and infections. You need a serosurvey, a known reporting share, or later data that shows the depletion of susceptibles.
- **5.** $R = 1 + 0.1 \times 8 = 1.8$.
- **6.** Standard deviations scale as $1/\sqrt{\lambda}$, so $\sqrt{10^6} = 1000$ times.

TRY-IT ANSWER

Try-it answer: at day 70 the cut starts after the peak, so the reported peak no longer changes with the start day and the local effect is near zero. The start day still matters a lot at other baselines: local results are local.

ANSWERS

Answers: Parts 3–4

- **7.** It acts almost entirely through interactions: alone it explains 1% of the variance, with its interactions 20%. It cannot be fixed safely.
- **8.** $N(k+2) = 4096 \times 12 = 49{,}152$ runs.
- **9.** It is a slope at one baseline with everything else fixed: steep at day 50, flat elsewhere, and blind to interactions.
- **10.** $(60-20) + \frac{2}{0.1}(70-60) = 40 + 200 = 240$.
- **11.** The same absolute error, since the medians match. A proper score usually prefers B; here 7.9 against 9.5.
- **12.** Planning surge capacity, where a shortage costs far more than spare beds: a model whose upper intervals are reliable may beat one with a smaller average error. Early warning of a new wave is another.

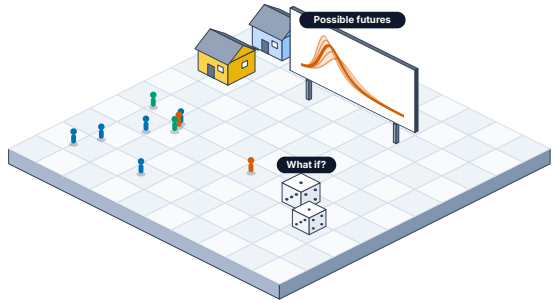
IN PLAIN WORDS

A model earns trust by saying how sure it is, showing what drives it, and being tested on the task it will be used for.

CREDITS

Sources and credits

- This deck accompanies Chapter 11 of T. Lazebnik, Epidemic Modeling: Principles, Algorithms, and Applications (Academic Press, 2026), the course textbook. It follows the chapter's topics in our own words; no text, figures, tables, exercises or code are copied from the book.
- The text, figures, examples and simulations were written, drawn and run for ACML; nothing is copied from the book. Numbers labelled "our run" come from our own code; counts from agencies and papers are credited where they appear.
- All numbers and curves come from our own made-up towns and outbreaks (production/explains/sims/epi_m11_sims.py), with values chosen to show mechanisms, not to describe any real disease, place or forecast.
- Named works, described in our own words: Raue et al., 2009 (profile likelihood); Wallinga and Lipsitch, 2007 (generation interval); Gutenkunst et al., 2007 (sloppiness); Morris, 1991; Jansen, 1999; Saltelli et al., 2010; Gneiting and Raftery, 2007; Bracher et al., 2021; Cramer et al., 2022.
- Companion videos: **Epidemic Modeling**, module 11, episodes 45–48. More from the lab: acml.teddylazebnik.com



Applied Computational Mathematics Lab

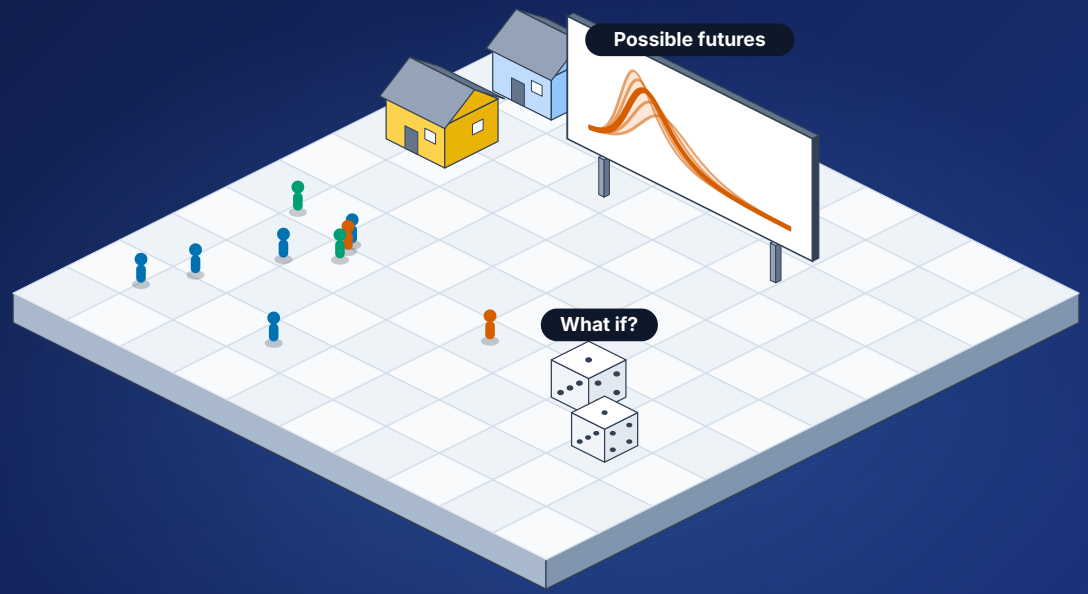
Module 11 complete

END OF MODULE 11

Uncertainty, sensitivity and model comparison

YOU COVERED

- Where uncertainty comes from
- Identifiability and confounding
- The sensitivity toolbox: Morris and Sobol
- Comparing models: scores and coverage



UP NEXT · MODULE 12

Computational considerations

Software, numerical stability, speed and scale

Based on Chapter 11 of T. Lazebnik, Epidemic Modeling: Principles, Algorithms, and Applications (Academic Press, 2026)