FRIEDRICH-ALEXANDER
UNIVERSITÄT
ERLANGEN-NÜRNBERG

FACULTY OF MEDICINE

Evaluating forecasts of infectious disease spread

Sebastian Meyer
Institute of Medical Informatics, Biometry, and Epidemiology
Friedrich-Alexander-Universität Erlangen-Nürnberg, Erlangen, Germany
21 March 2019

Based on joint work with Leonhard Held (University of Zurich):

Held and Meyer (2019). Forecasting Based on Surveillance Data. In: *Handbook of Infectious Disease Data Analysis*. Chapman & Hall/CRC. arXiv:1809.03735

Epidemics are hard to predict

World Health Organization (2014)

> *Forecasting disease outbreaks is still in its infancy, however, unlike weather forecasting, where substantial progress has been made in recent years.*

Epidemics are hard to predict

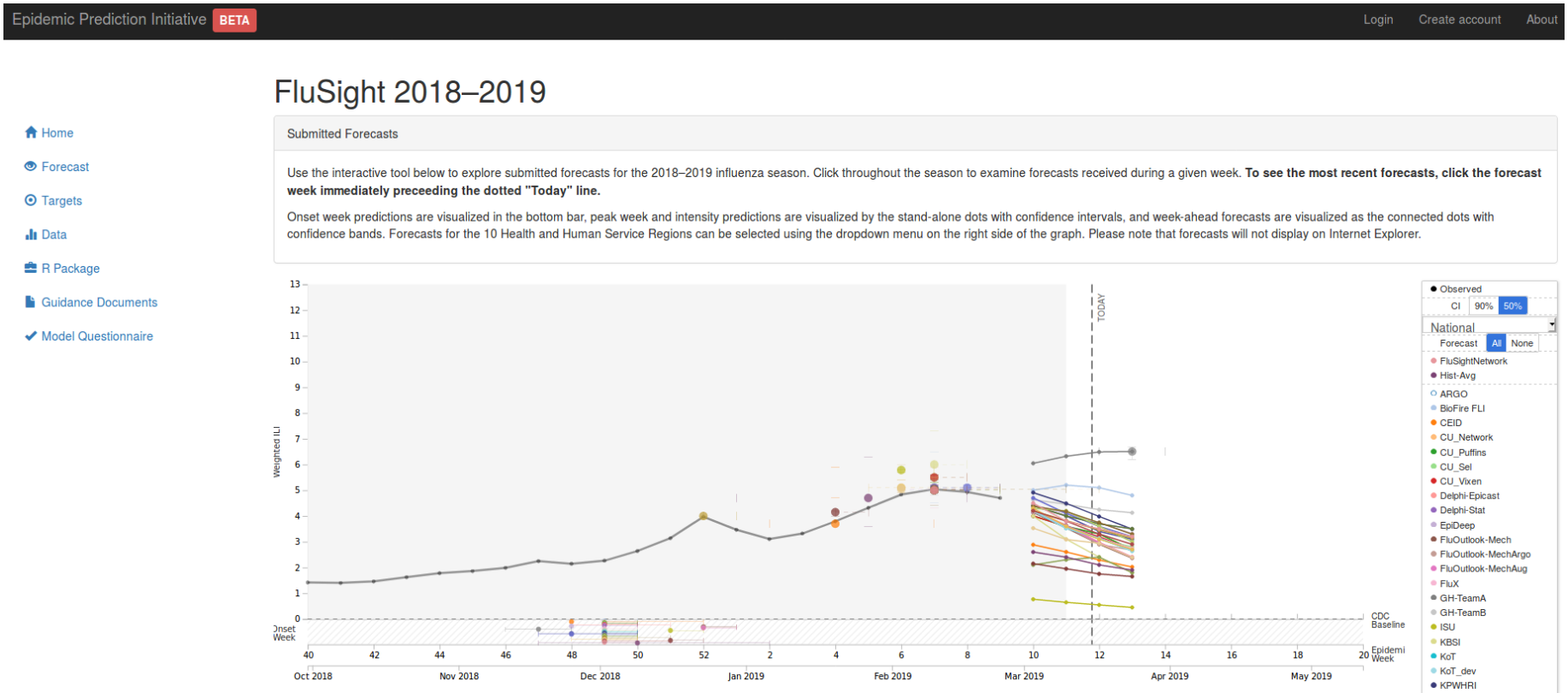
World Health Organization (2014)

> *Forecasting disease outbreaks is still in its infancy, however, unlike weather forecasting, where substantial progress has been made in recent years.*

Meanwhile . . .

- **Epidemic Prediction Initiative** (Centers for Disease Control and Prevention, 2016): online platform collecting real-time forecasts by various research groups
- Adoption of **forecast assessment techniques** from weather forecasting (Held, Meyer, & Bracher, 2017)
- Integration of **social contact patterns** (Meyer & Held, 2017), **human mobility data** (Pei, Kandula, Yang, & Shaman, 2018), and **internet data** (Osthus, Daughton, & Priedhorsky, 2019)

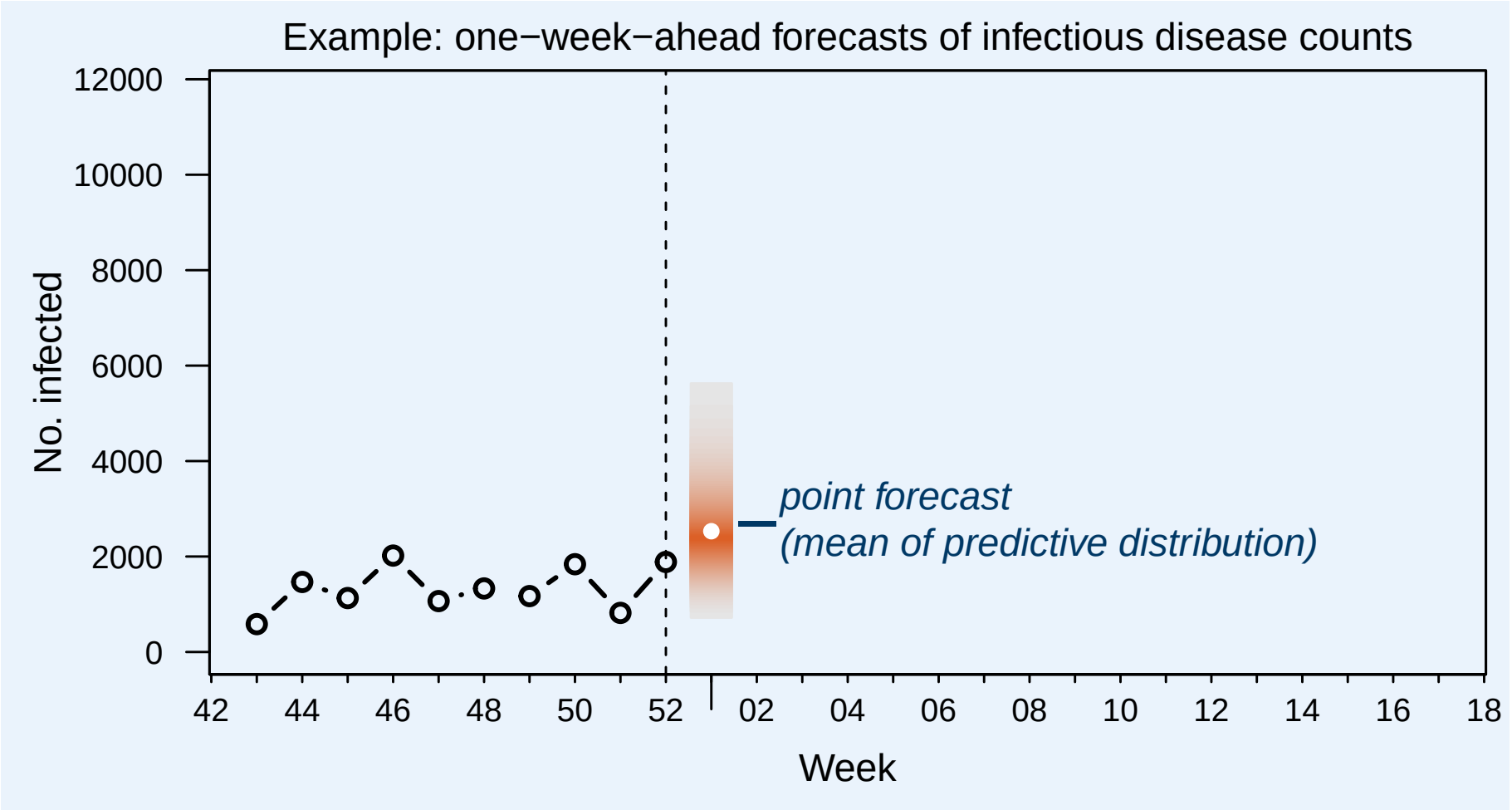
CDC FluSight challenge (https://predict.cdc.gov/)



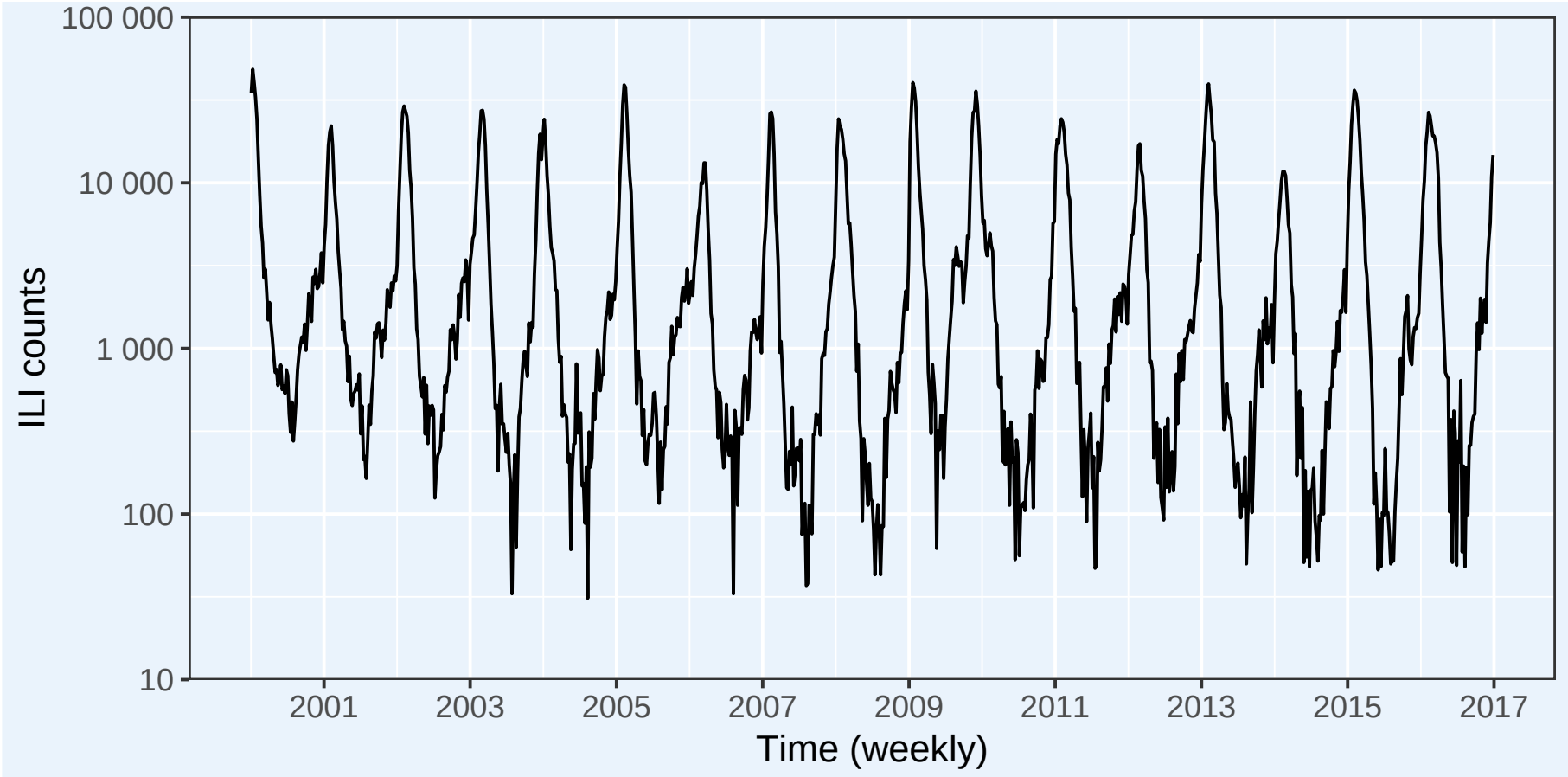
Multiple **forecasting targets** for influenza-like illness (ILI):

- short-term doctor visits: 1 to 4 weeks ahead
- seasonal targets: onset week, peak week, peak incidence

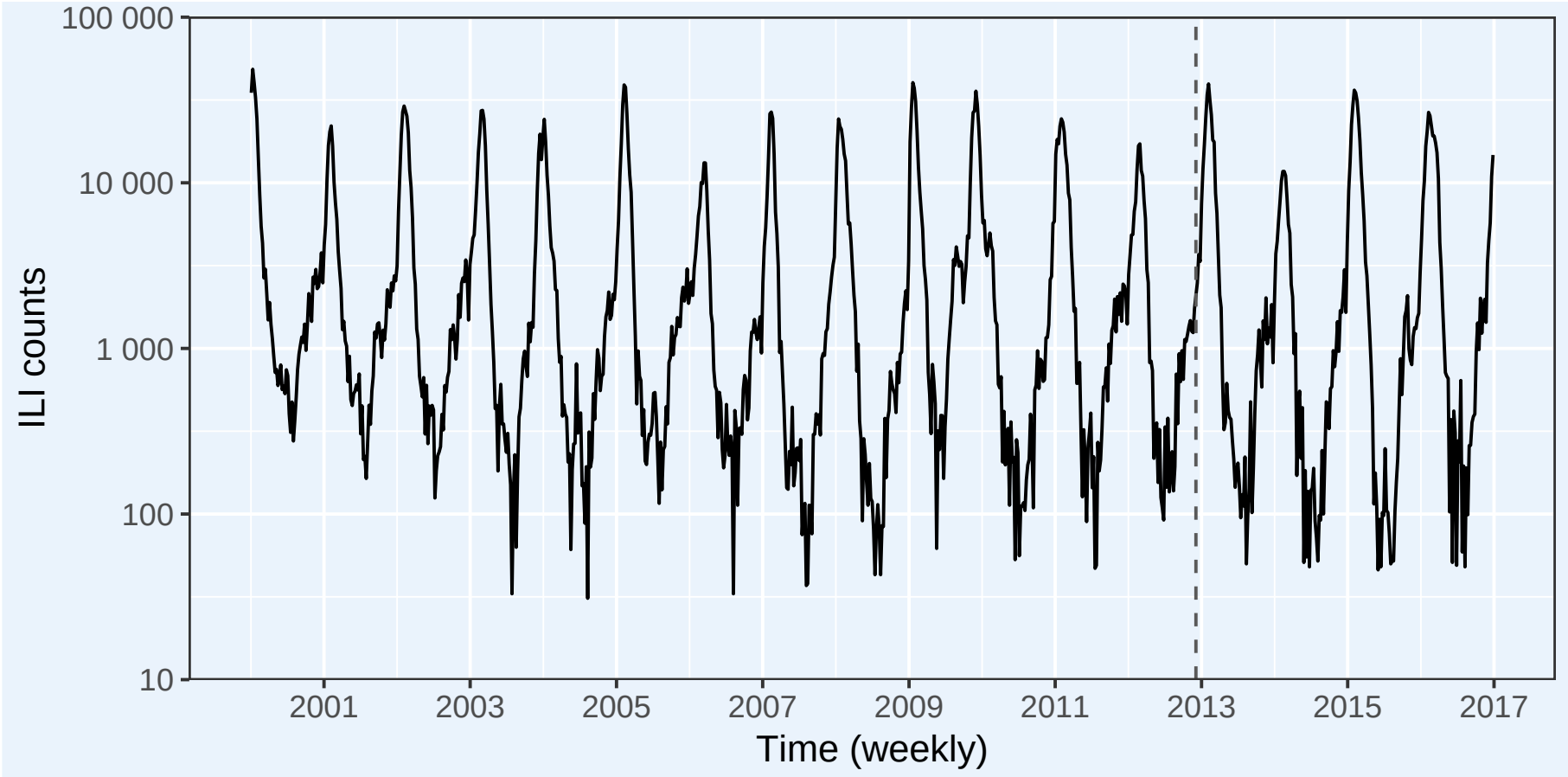
“Forecasts should be probabilistic” (Gneiting & Katzfuss, 2014)



Case study: Weekly ILI counts in Switzerland, 2000–2016

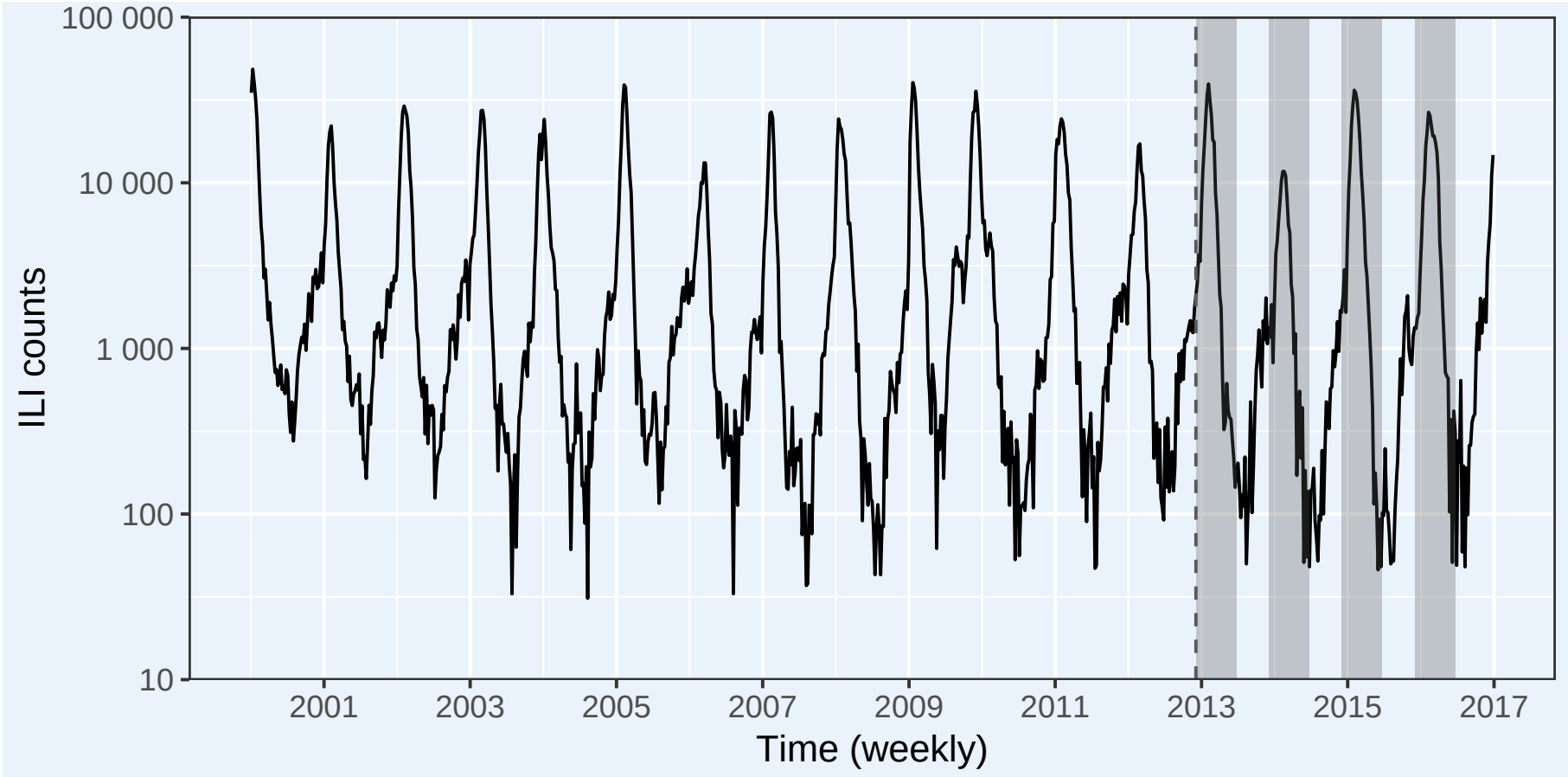


Case study: Weekly ILI counts in Switzerland, 2000–2016



1. Rolling one-week-ahead forecasts in the test period (from December 2012)

Case study: Weekly ILI counts in Switzerland, 2000–2016



1. Rolling one-week-ahead forecasts in the test period (from December 2012)
2. Seasonal forecasts of the epidemic curve (30-weeks-ahead from December)

Evaluating forecasts

- **Goal**: compare predictive performance of different models
  1. We evaluate point forecasts by RMSE or MAE,

     not correlation between point predictions and observations
  2. We assess the whole distribution of probabilistic forecasts

Evaluating forecasts

- **Goal**: compare predictive performance of different models
  1. We evaluate point forecasts by RMSE or MAE,

     not correlation between point predictions and observations
  2. We assess the whole distribution of probabilistic forecasts
- **Paradigm**: maximize sharpness subject to calibration
  - Calibration: statistical consistency of forecast F and observation y
  - Sharpness: width of prediction intervals

Evaluating forecasts

- **Goal**: compare predictive performance of different models
  1. We evaluate point forecasts by RMSE or MAE,

     not correlation between point predictions and observations
  2. We assess the whole distribution of probabilistic forecasts
- **Paradigm**: maximize sharpness subject to calibration
  - Calibration: statistical consistency of forecast F and observation y
  - Sharpness: width of prediction intervals
- **Assessment techniques**:
  - Histogram of $PIT = F(y)$ values to informally check calibration
  - Proper scoring rules $S(F, y)$ as summary measures of predictive performance addressing both calibration and sharpness

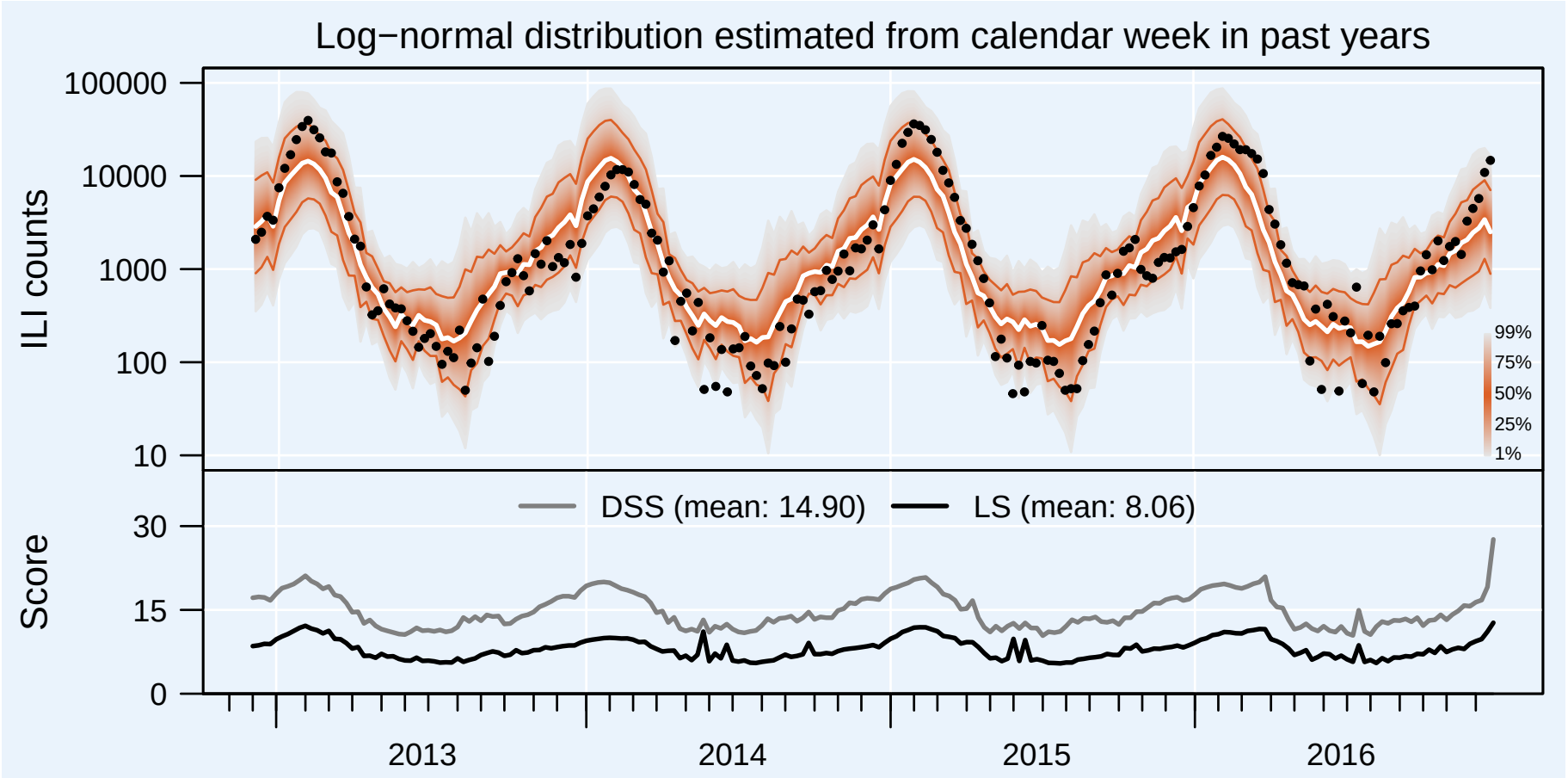
Proper scoring rules

- Scoring rule $S(F, y)$ quantifies discrepancy between forecast and observation

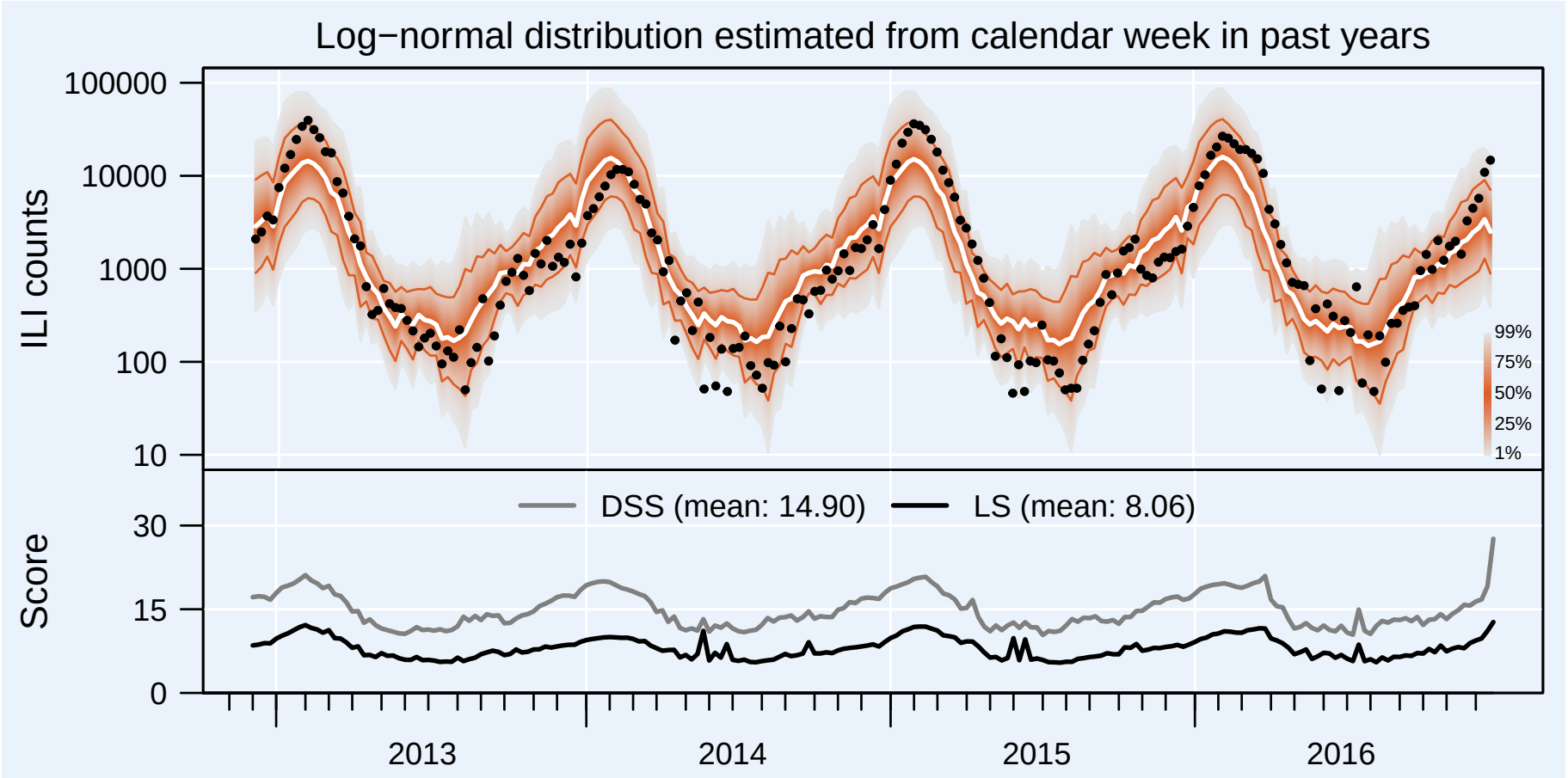
  $\rightarrow$ something we would like to *minimize*

- Propriety: forecasting with the true distribution is optimal
- Simple example: squared error score $SES(F, y) = (y - \mu_F)^2$
- Compute average score over a test set of forecasts, e.g., (R)MSE
- We will use the following scoring rules:
  - Logarithmic score: $LS(F, y) = -\log f(y)$
  - Dawid-Sebastiani score: $DSS(F, y) = \log(\sigma_F^2) + \frac{(y-\mu_F)^2}{\sigma_F^2}$

"Naive" forecast stratified by calendar week

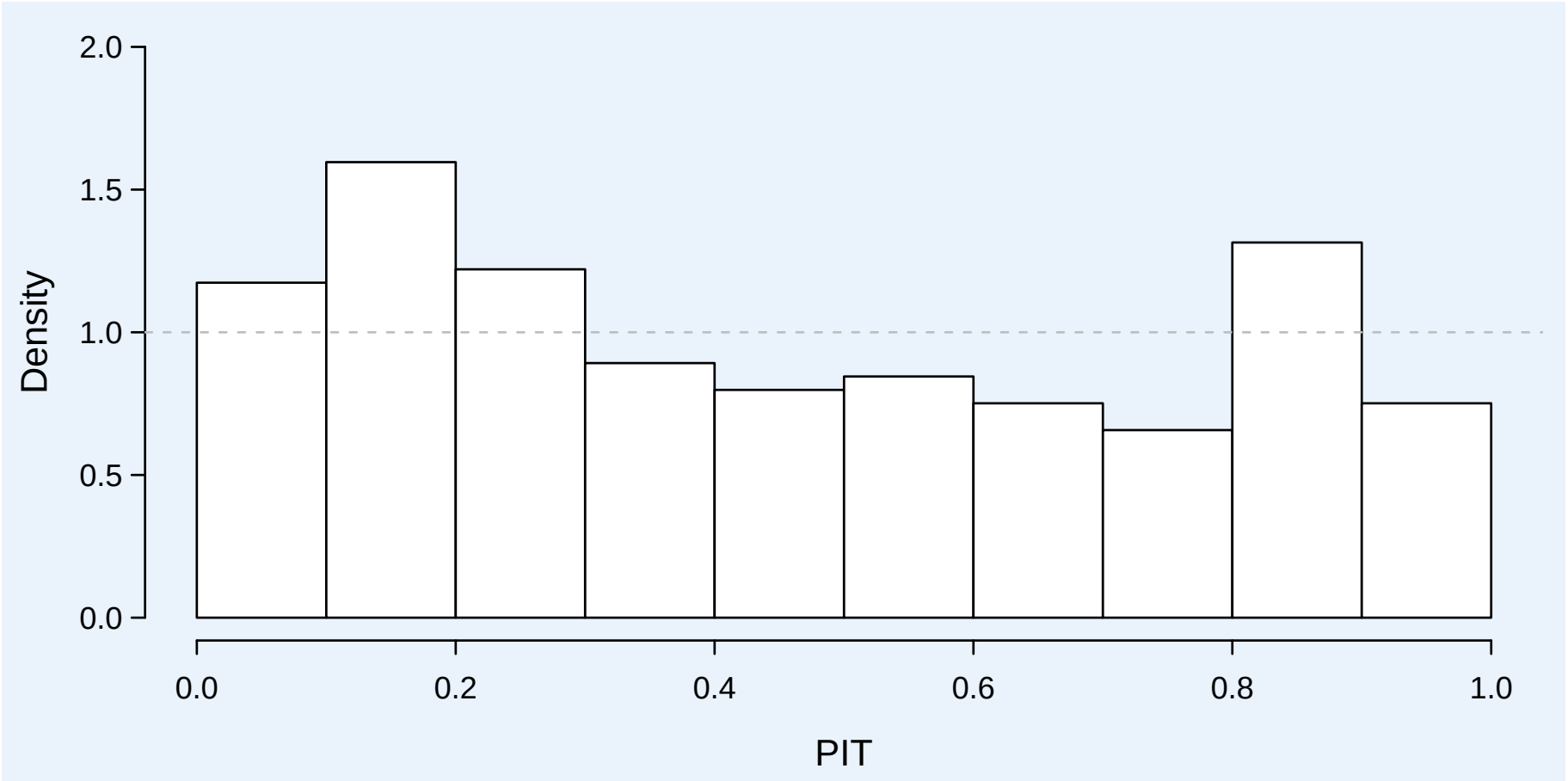


"Naive" forecast stratified by calendar week

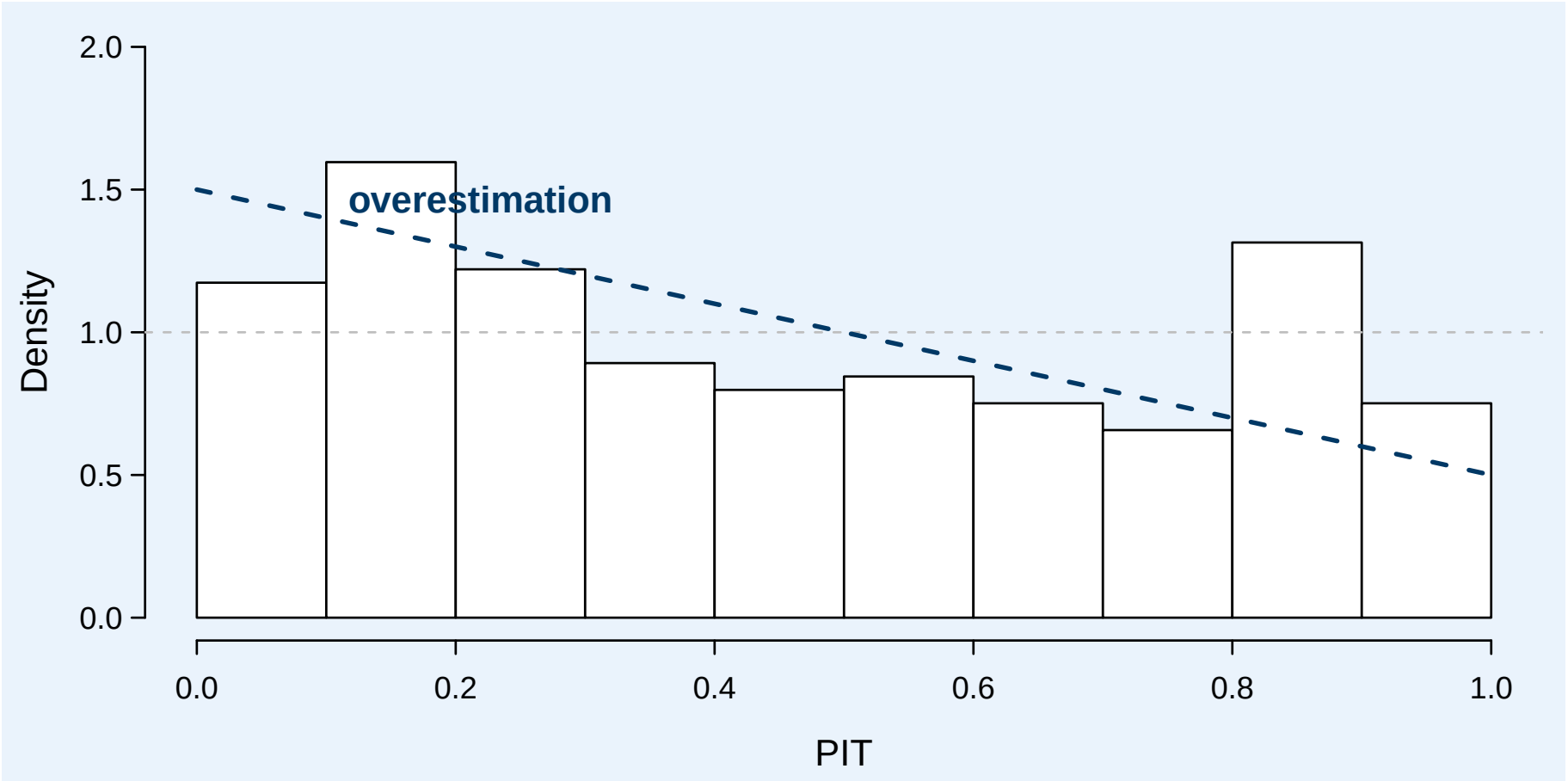


- Wide prediction intervals, RMSE = 5010 cases
- Well calibrated? PIT histogram summarizes location of observations in the fan

PIT histogram of the 213 one-week-ahead forecasts

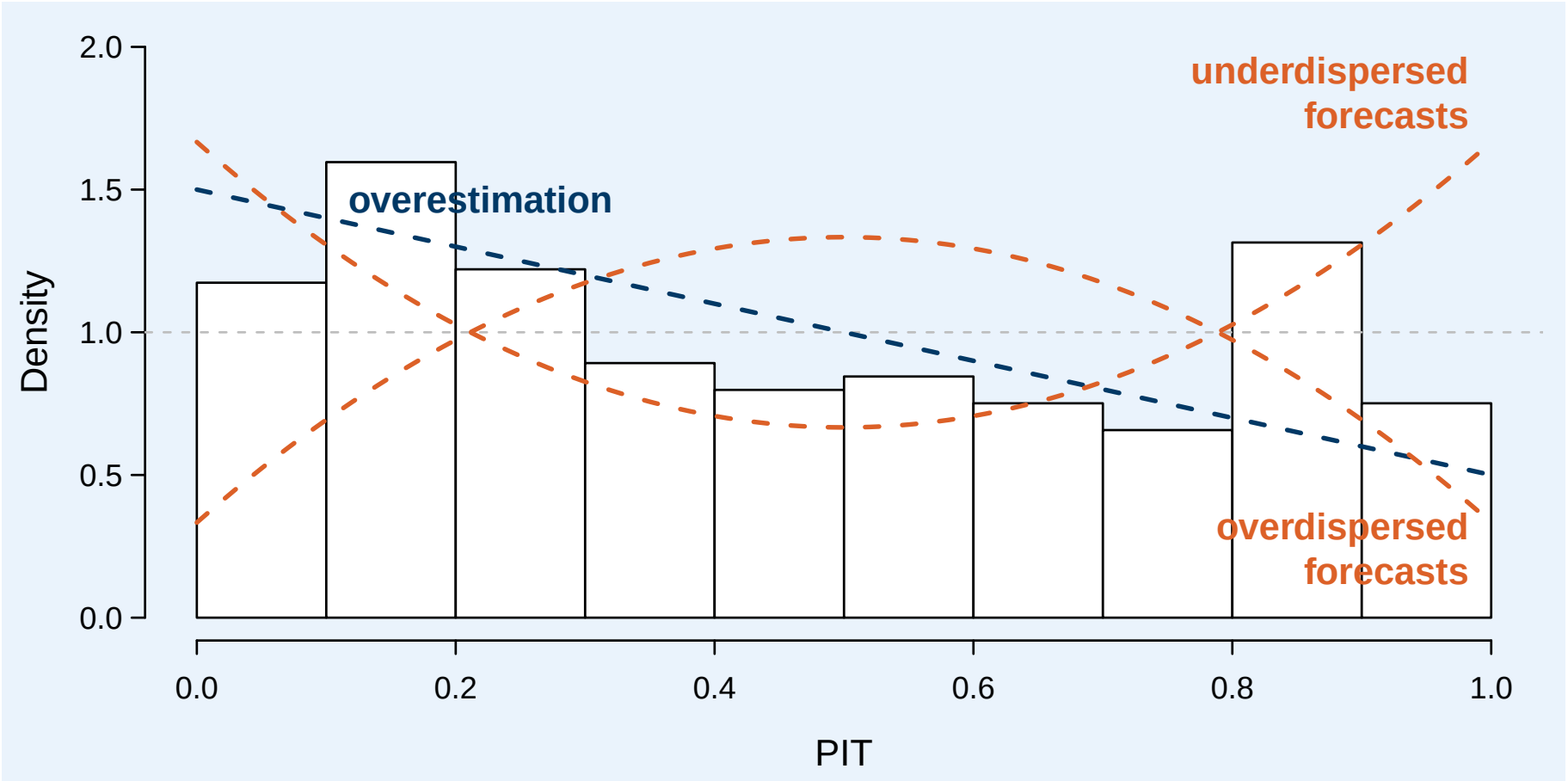


PIT histogram of the 213 one-week-ahead forecasts



- Counts tend to be lower than predicted

PIT histogram of the 213 one-week-ahead forecasts



- Counts tend to be lower than predicted
- No clear-cut evidence of miscalibration

Useful *statistical models* to forecast epidemic spread

- Can we do better with more sophisticated time series models?
- Scope: *well-documented open-source* R implementations
- We compare four different models:
  - `forecast::auto.arima()` for log-counts $\rightarrow$ ARMA(2,2)
  - `glarma::glarma()` $\rightarrow$ NegBin-ARMA(4,4)
  - `surveillance::hhh4()`: "endemic-epidemic" NegBin model
  - `prophet::prophet()` for log-counts: linear regression model
- All models account for yearly seasonality and a Christmas effect

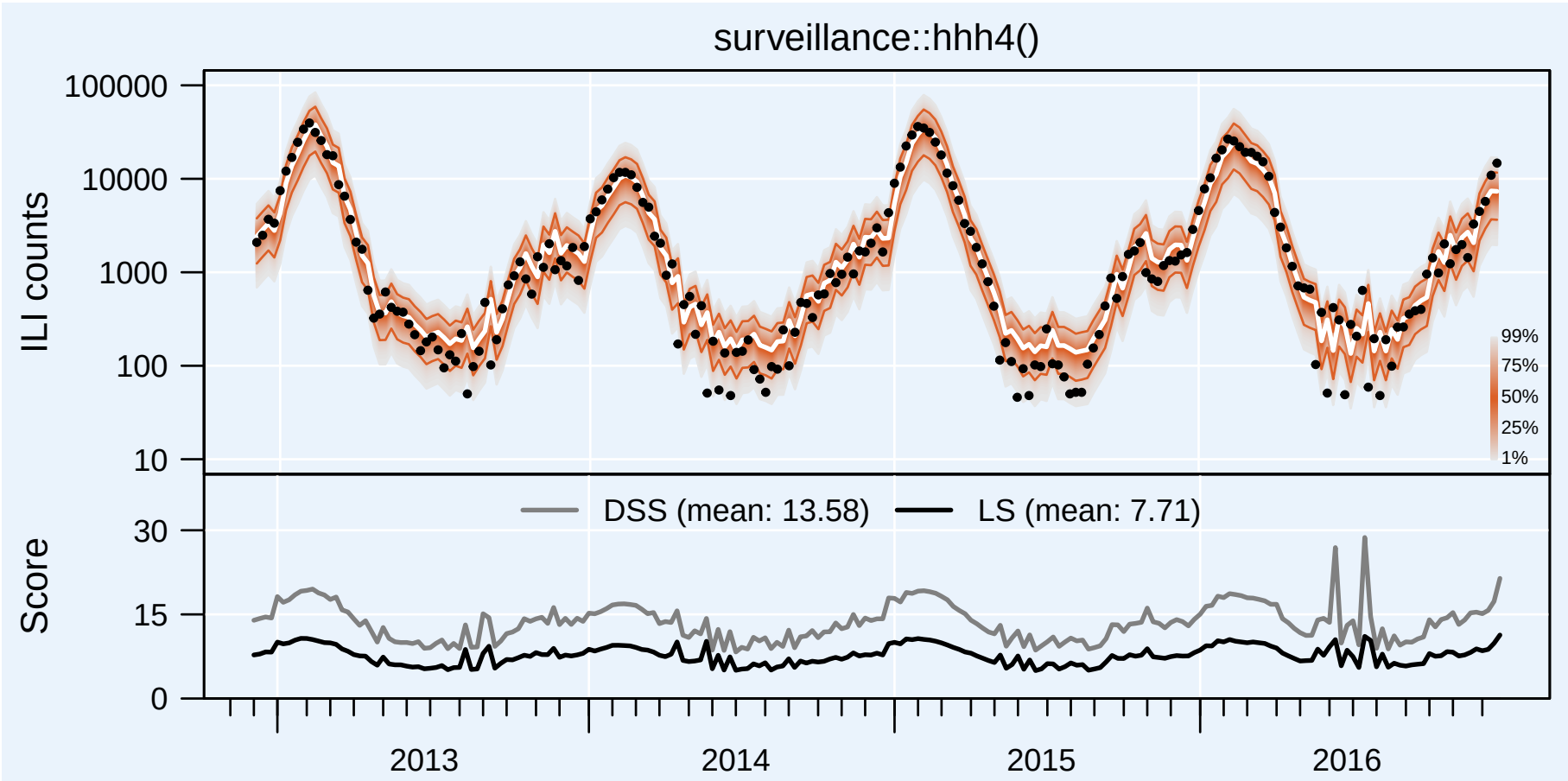
Performance of rolling one-week-ahead forecasts

Average scores and runtime based on 213 one-week-ahead forecasts:

Method	RMSE	DSS	LS	runtime [s]
`arima`	2287	13.78	7.73	0.51
`glarma`	2450	13.59	7.71	1.49
`hhh4`	**1769**	**13.58**	**7.71**	0.02
`prophet`	5614	15.00	8.03	3.01
naive	5010	14.90	8.06	**0.00**

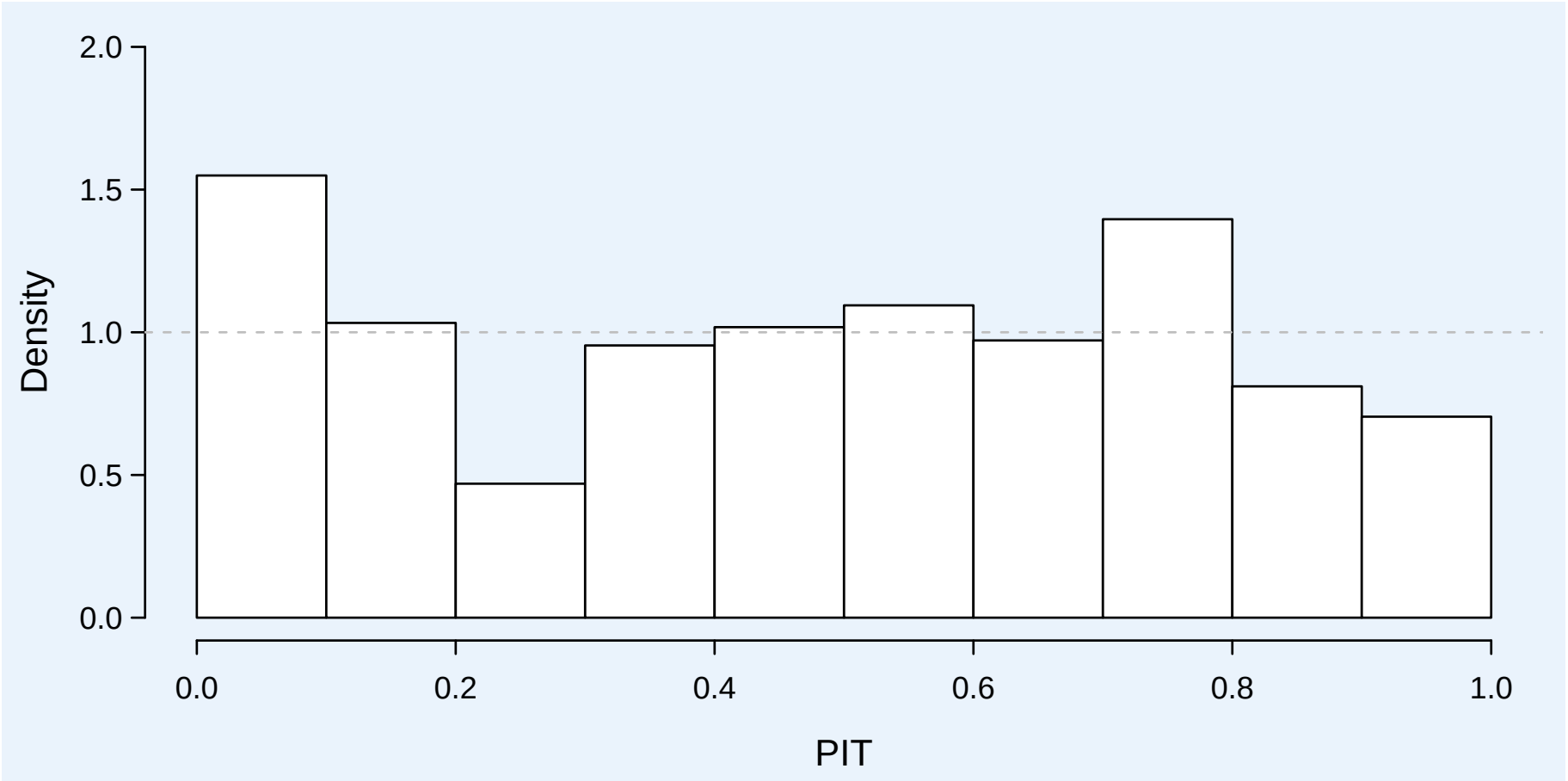
- All methods are reasonably fast
- The two NegBin models score best
- `prophet` does not outperform the naive approach

hhh4-based one-week-ahead forecasts



- Sharper than naive forecast (drawback in wiggly off-season 2016)
- Seaonsal autoregressive effect adapts to yearly peaks

PIT histogram for hhh4-based one-week-ahead forecasts



- Calibration similar to naive forecasts
- Off-season counts in lower tail of forecast distribution

Performance of seasonal forecasts of the epidemic curve

Average scores and runtime based on four 30-weeks-ahead forecasts:

Method	RMSE	DSS	LS	runtime [s]
`arima`	8471	16.43	8.88	0.48
`glarma`	**5558**	19.61	9.12	4.13
`hhh4`	8749	16.13	9.25	0.46
`prophet`	7627	16.44	8.91	0.92
naive	6527	**15.99**	**8.86**	**0.00**

- None of the sophisticated models outperforms the naive approach
- DSS and LS rank the models differently
- Large DSS for `glarma` is due to large uncertainty
- Forecasting the epidemic curve right from the season start is truly ambitious

Discussion

- Key requirements to forecast infectious disease incidence:
  1. **Routine public health surveillance data** (notifiable diseases)
  2. **Forecasting targets** and **evaluation methods**
  3. Useful **statistical models** to forecast epidemic spread
- The case study exemplified the necessary steps
- Data and reproduction code: https://HIDDA.github.io/forecasting/
- Of course, different rankings might result with other time series

Discussion

- Key requirements to forecast infectious disease incidence:
  1. **Routine public health surveillance data** (notifiable diseases)
  2. **Forecasting targets** and **evaluation methods**
  3. Useful **statistical models** to forecast epidemic spread
- The case study exemplified the necessary steps
- Data and reproduction code: https://HIDDA.github.io/forecasting/
- Of course, different rankings might result with other time series

- Combination of different forecasting methods (ensemble forecasts)
- Extension of models and evaluation techniques for multivariate **forecasts by region or age group** $\rightarrow$ incorporate travel and social contact patterns
- Incorporation of reporting delays and underreporting

References

Centers for Disease Control and Prevention. (2016). Flu activity forecasting website launched.

Gneiting, T., & Katzfuss, M. (2014). Probabilistic forecasting. *Annual Review of Statistics and Its Application*, *1*(1), 125–151. https://doi.org/10.1146/annurev-statistics-062713-085831

Held, L., & Meyer, S. (2019). Forecasting based on surveillance data. In L. Held, N. Hens, P. D. O'Neill, & J. Wallinga (Eds.), *Handbook of infectious disease data analysis*. Chapman & Hall/CRC.

Held, L., Meyer, S., & Bracher, J. (2017). Probabilistic forecasting in infectious disease epidemiology: The 13th Armitage lecture. *Statistics in Medicine*, *36*(22), 3443–3460. https://doi.org/10.1002/sim.7363

Meyer, S., & Held, L. (2017). Incorporating social contact data in spatio-temporal models for infectious disease spread. *Biostatistics*, *18*(2), 338–351. https://doi.org/10.1093/biostatistics/kxw051

Osthus, D., Daughton, A. R., & Priedhorsky, R. (2019). Even a good influenza forecasting model can benefit from internet-based nowcasts, but those benefits are limited. *PLOS Computational Biology*, *15*(2), 1–19. https://doi.org/10.1371/journal.pcbi.1006599

Pei, S., Kandula, S., Yang, W., & Shaman, J. (2018). Forecasting the spatial transmission of influenza in the United States. *Proceedings of the National Academy of Sciences of the United States of America*, *115*(11), 2752–2757. https://doi.org/10.1073/pnas.1708856115

World Health Organization. (2014). Anticipating epidemics. *Weekly Epidemiological Record*, *89*(22), 244. Retrieved from http://www.who.int/wer