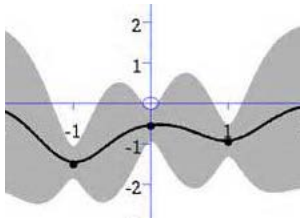
Improved Inference by accounting for Model Discrepancy

Thomas Wutzler
MPI-Biogeochemistry, Jena


Thomas Wutzler

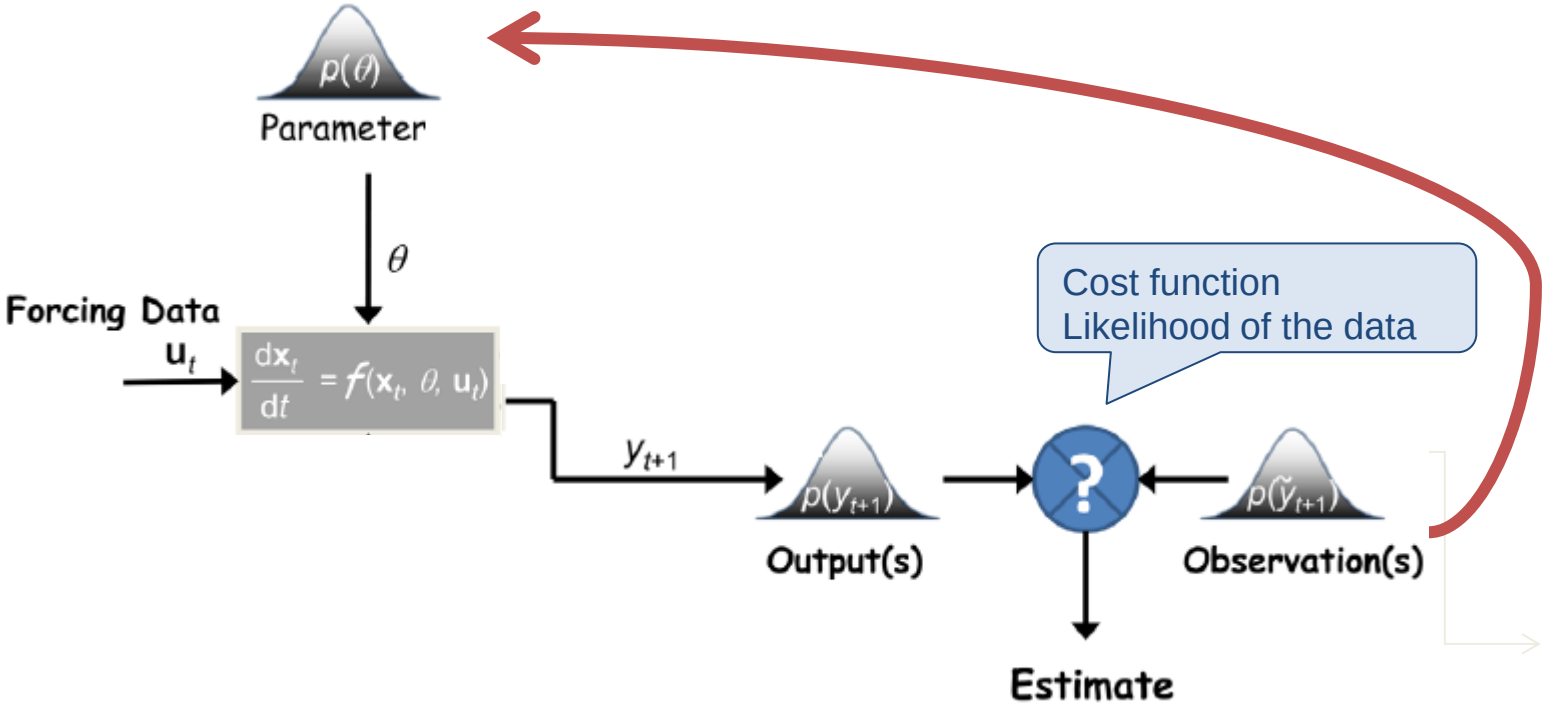




Gaussian Process (GP)

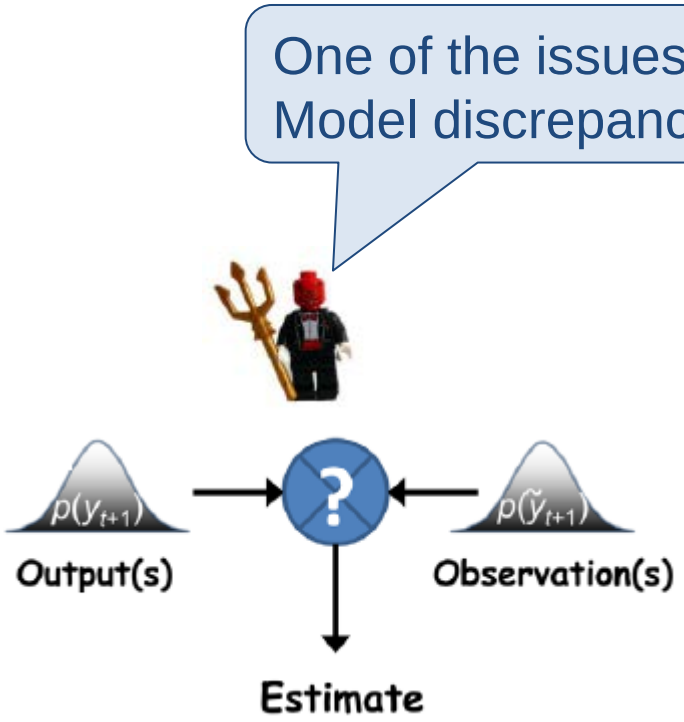
observation = model + discrepancy + noise

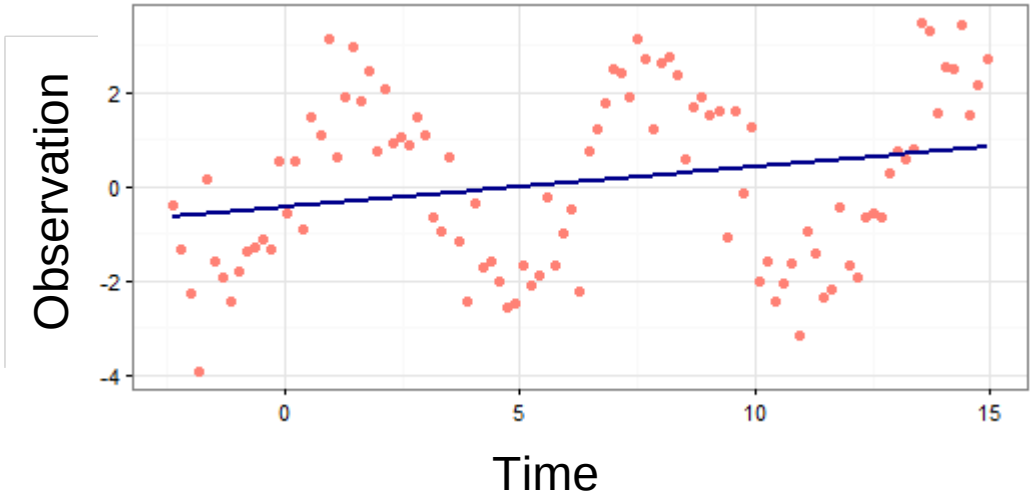
Model inversion:
Inferring uncertainty of model parameters



adapted from Vrugt 2013

Model discrepancy
violates assumption of uncorrelated errors

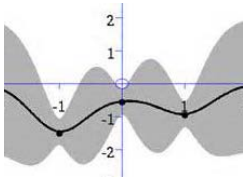




adapted from Vrugt 2013

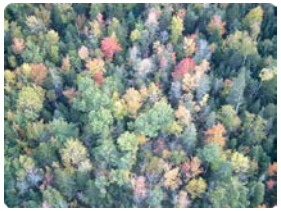
What you will hear

Gaussian Processes (GP) for discrepancy



Basic example

Innovations for large data streams



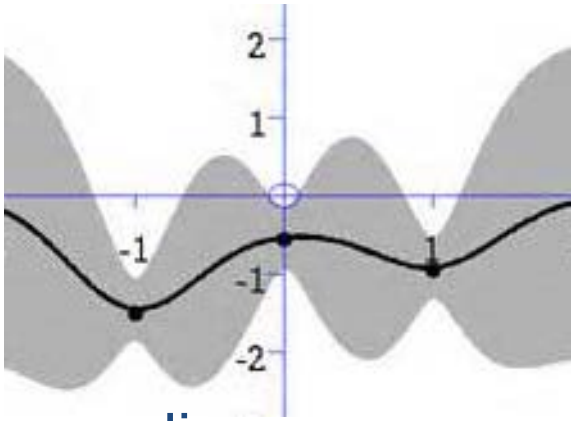
Proof of concept real world example

Gaussian Processes (GP) can account for model discrepancy

Gaussian Process (GP)

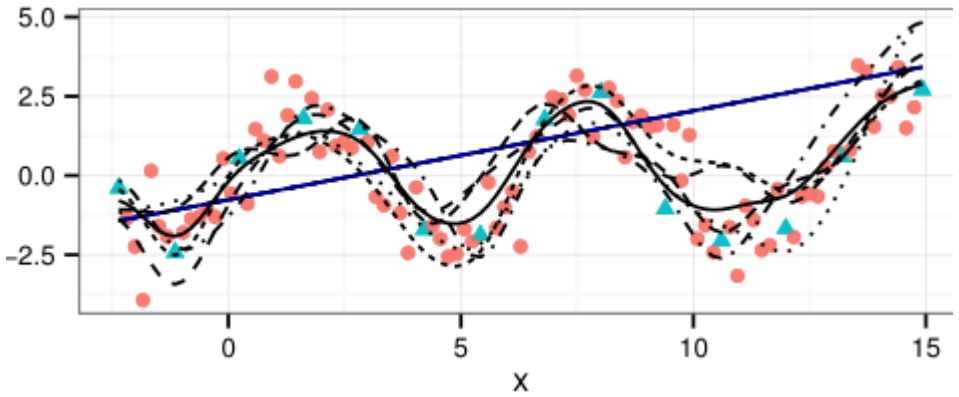
=

+



+

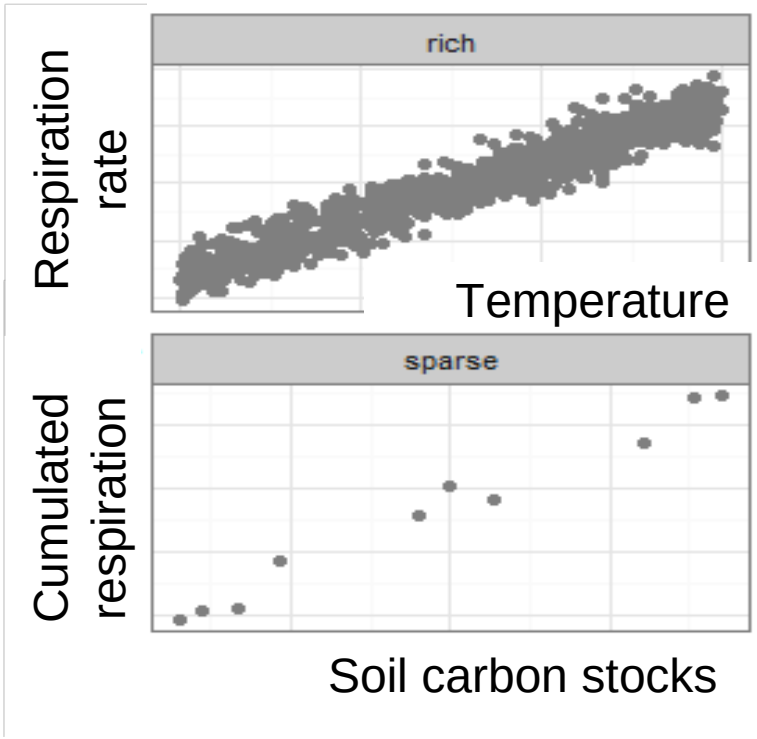
observation = model + discrepancy + noise



$$\mathbf{o} = \mathbf{g}(\boldsymbol{\theta}) + \boldsymbol{\delta} + \boldsymbol{\epsilon}$$
$$\boldsymbol{\epsilon} \sim N(0, \sigma_\epsilon^2)$$
$$\boldsymbol{\delta} \sim GP(\mathbf{0}, \psi, \sigma_d^2).$$

Kennedy and O'Hagan (2001); Brynjarsdottir and O'Hagan, (2014); Rasmussen and Williams (2006)

Basic synthetic example
clarifies effects of GP and imbalanced streams



Soil respiration measurements

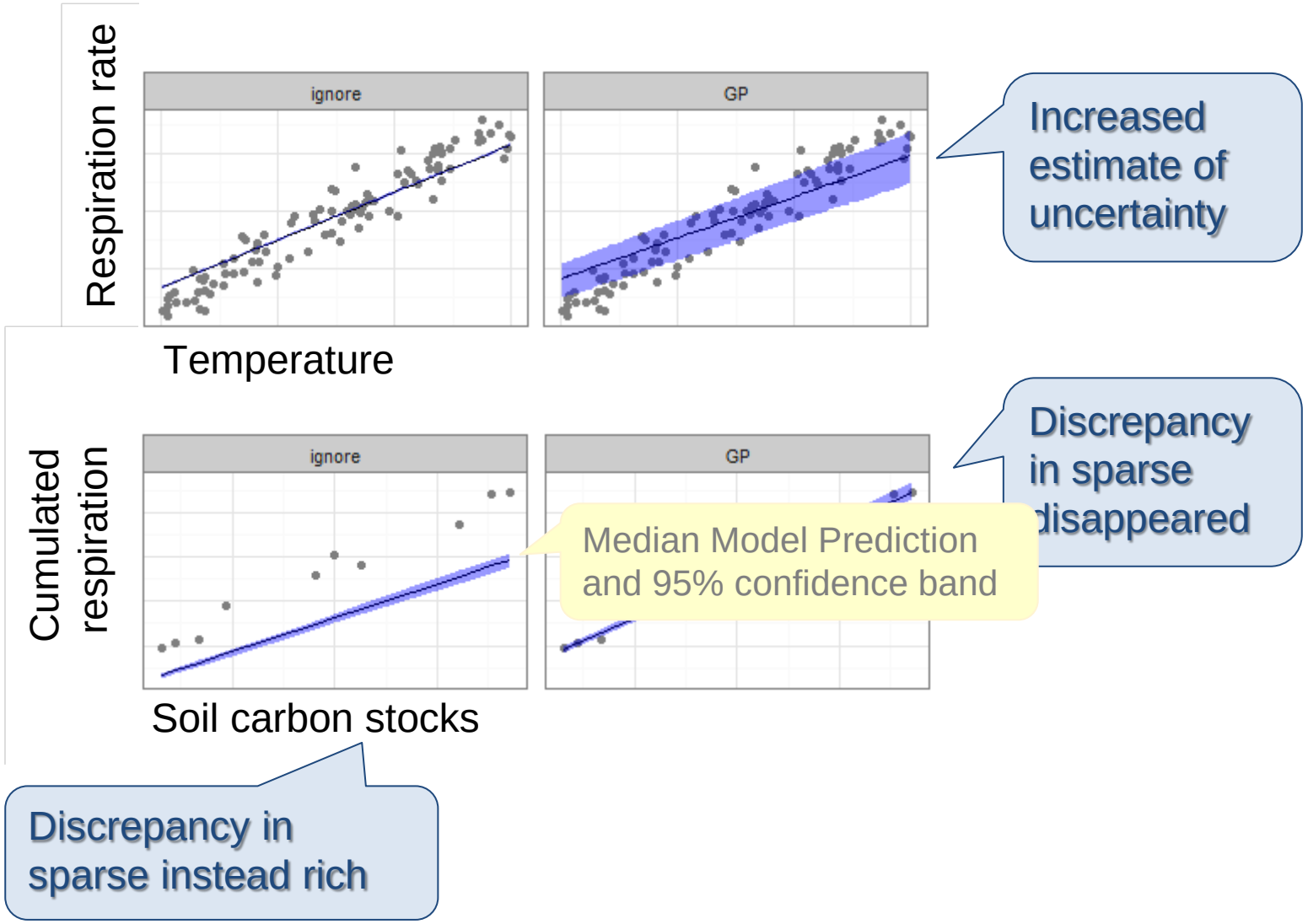
Rich: fast resonses
respiration rate ~ T
(n=1000)

Sparse: long term changes
cumulated respiration ~ soil carbon
(n=10)

Generated artificial observation
simple linear model + noise

Bayesian model inversion (MCMC)
using model with **bias** in
prediction of **rich** stream

GP correctly
locates model deficiency in sparse stream



Developments for large data
Treating discrepancy as hidden variable

State of the art

- Sampling model discrepancies
  - Low acceptance rate and slow mixing
- Fixed supporting points
  - Unrealistically high likelihood at lucky choices

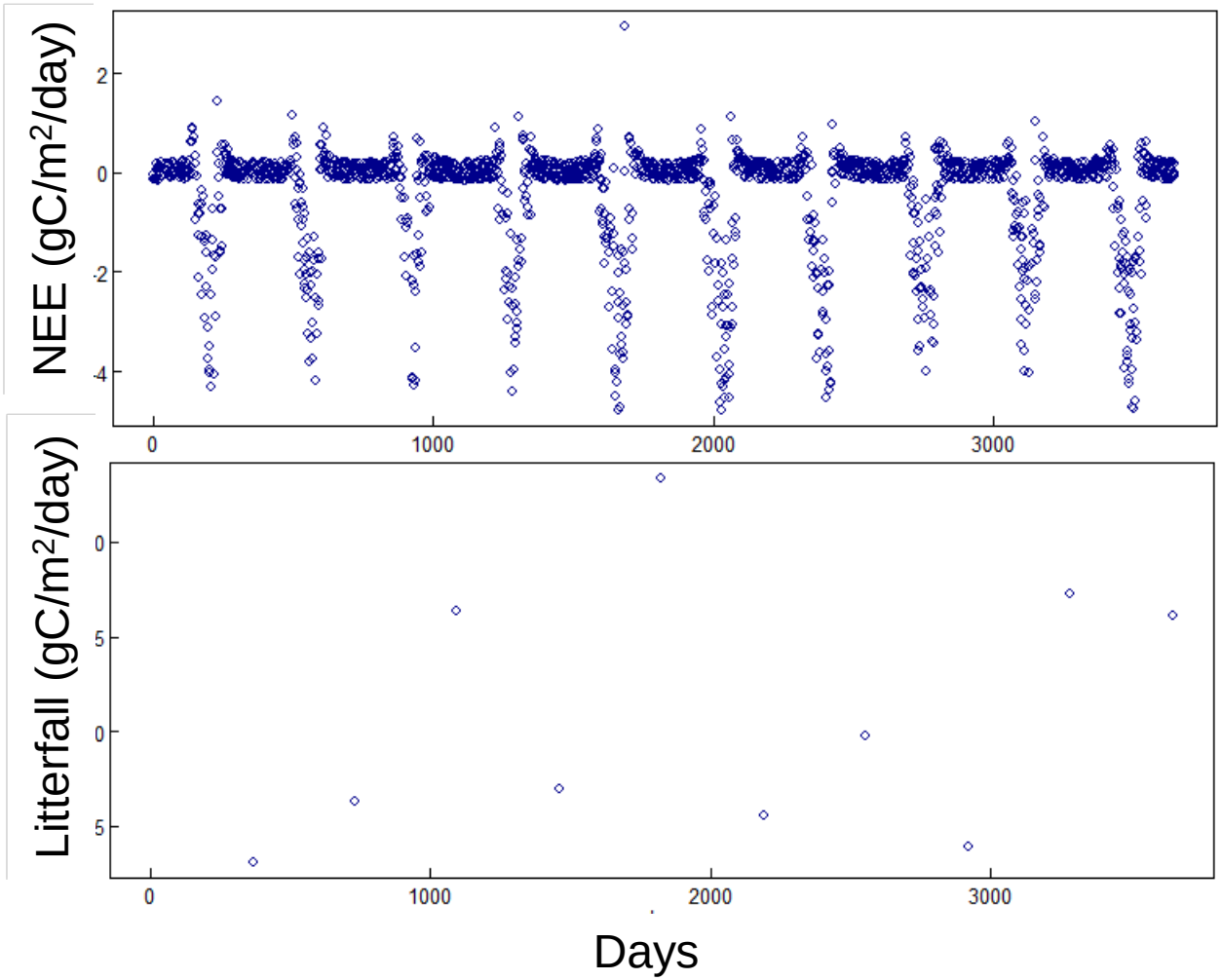
Innovation

- Computing expected model discrepancies for each new draw of parameters
- Respacing supporting points based on correlation length
- Randomness in supporting points

Brynjarsdottir and O'Hagan, (2014); Wutzler (in prep.) Inverse problems

Proof of concept

DALEC model inverted at Howland forest



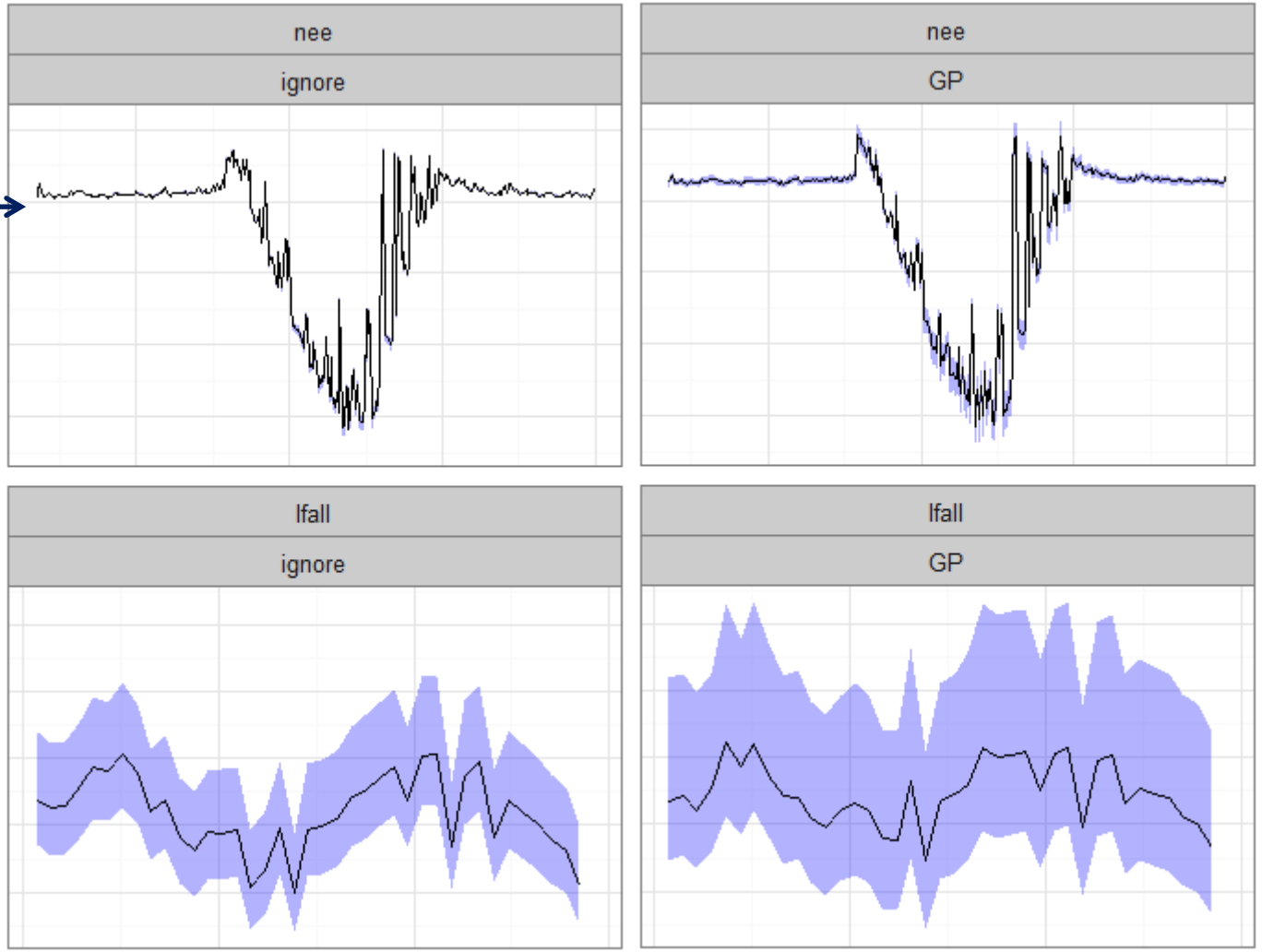
Howland Forest (whrc.org)

DALEC: Williams et al. 2005

Successful inversion

Increased uncertainty estimate

over-confidende in rich data streams



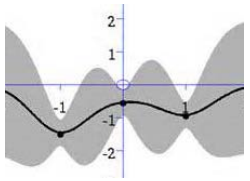
40 Years litter fall, 40th year of NEE and respiration

Block approach: Wutzler 2014

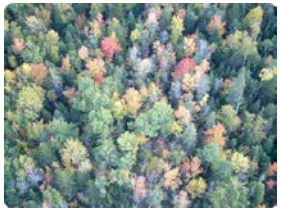
Conclusion
improved model inference using GP

- We need to *account for model discrepancy* especially with imbalanced data streams

- Discrepancy can be *modeled by Gaussian Processes*



- Several *methodological innovations* allow application to real world rich data stream



Thanks to
- Nuno Carvalhais and Paul Bodesheim for fruitful discussions
- T. Keenan and D. Hollinger for permission to use the Howland data and model setup
- German research council (DFG) for funding within AquaDiva collaborative research center 1076

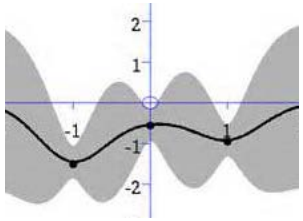
Paper, that this talk is based on:

Wutzler T (in prep) Multiple constraints in inverse problems: The importance of model discrepancy. *Inverse Problems journal*

```
install.packages(      "blockDemac",
    repos=c("http://R-Forge.R-project.org","@CRAN@"), type="source")
vignette("TwoDenGP")
```

Main Message:

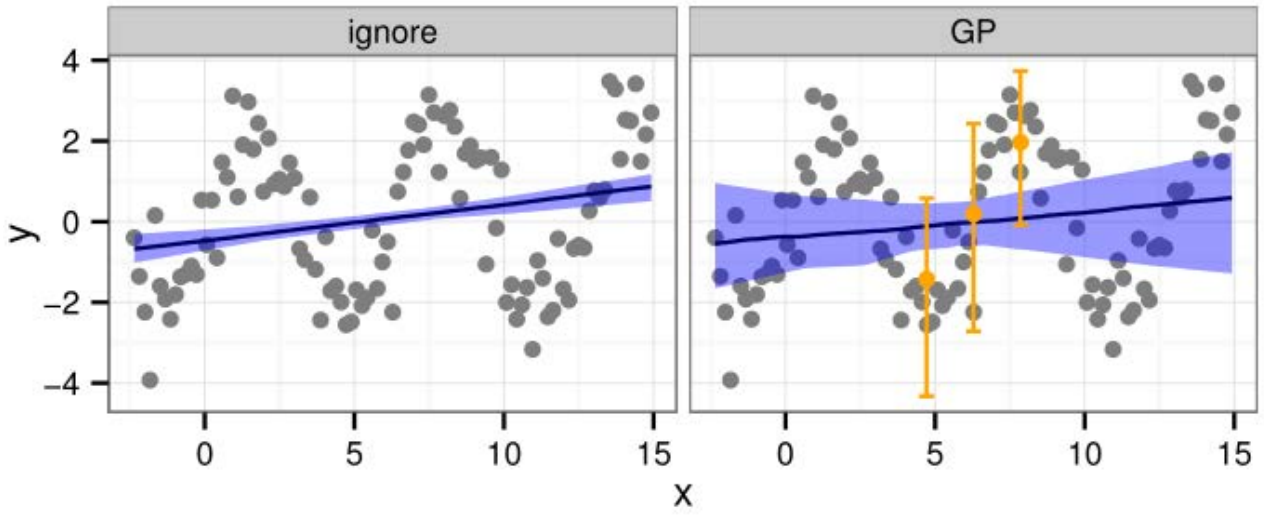
Gaussian Process (GP)



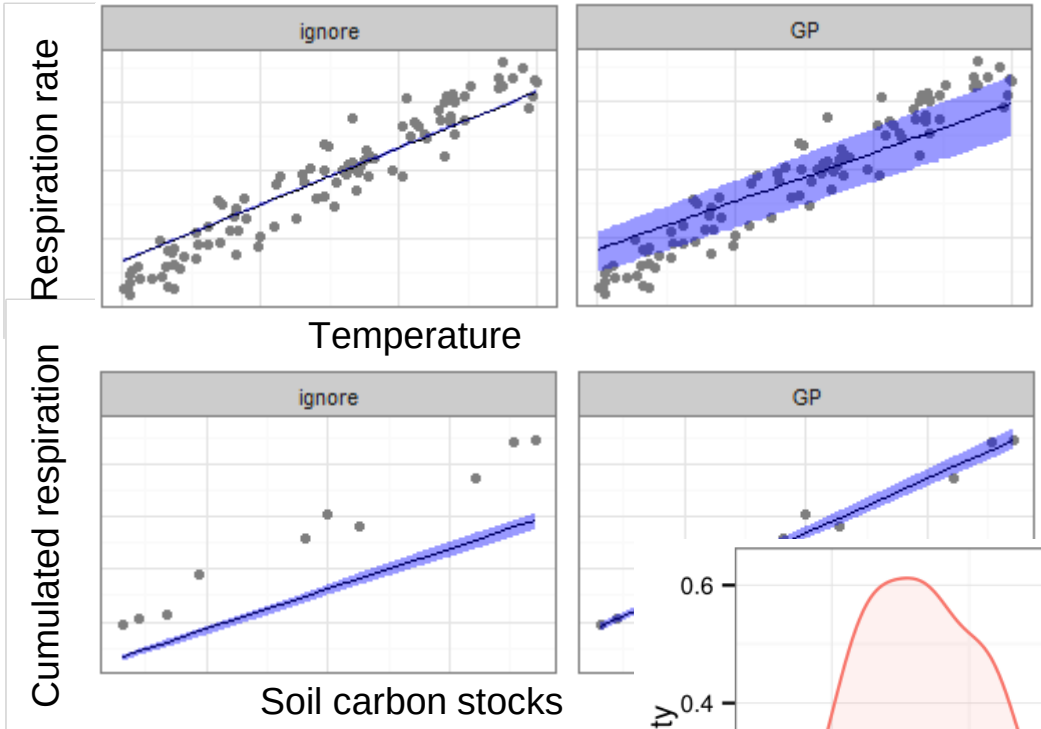
model discrepancy

Thomas Wutzler
twutz@bgc-jena.mpg.de

Estimate of uncertainty increases



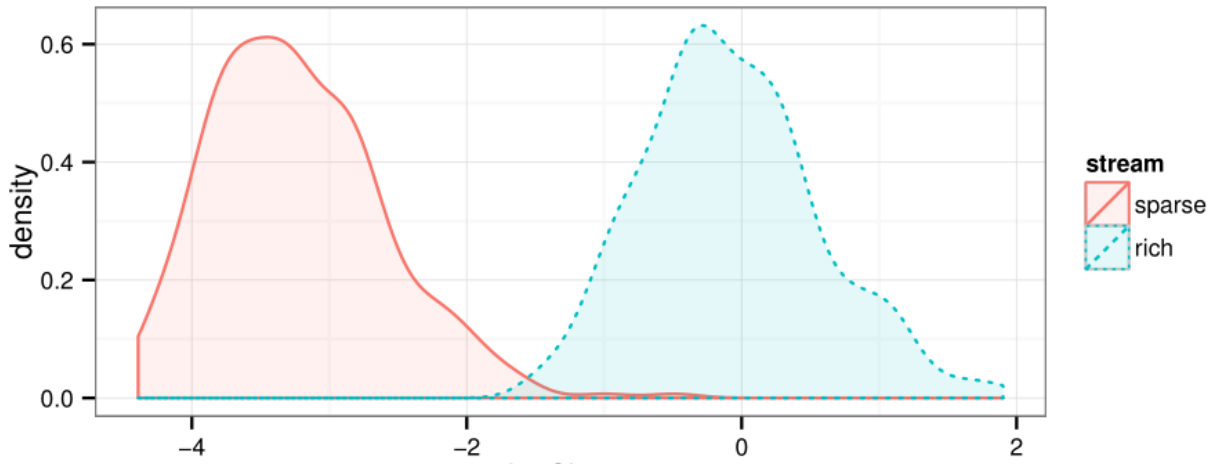
GP correctly
locates model deficiency in sparse stream



Increased estimate of uncertainty

Discrepancy in sparse disappeared

Discrepancy in sparse instead rich



Model discrepancy can be modelled as hidden variable

$$p(\boldsymbol{\theta}|\mathbf{o}) = \frac{p\left(\boldsymbol{\theta}, \hat{\boldsymbol{\delta}}|\mathbf{o}\right)}{p\left(\hat{\boldsymbol{\delta}}|\boldsymbol{\theta}, \mathbf{o}\right)} \propto \frac{p(\mathbf{o}|\boldsymbol{\theta}, \hat{\boldsymbol{\delta}})\, p(\hat{\boldsymbol{\delta}}|\boldsymbol{\theta})\, p(\boldsymbol{\theta})}{p\left(\hat{\boldsymbol{\delta}}|\boldsymbol{\theta}, \mathbf{o}\right)} \tag{A.12}$$

$$= \prod_k p(\mathbf{o}_k|\boldsymbol{\theta}, \hat{\boldsymbol{\delta}}_k) \frac{p(\hat{\boldsymbol{\delta}}_k|\boldsymbol{\theta})}{p\left(\hat{\boldsymbol{\delta}}_k|\boldsymbol{\theta}, \mathbf{o}_k\right)} p(\boldsymbol{\theta}) \tag{A.13}$$

$$\frac{p(\hat{\boldsymbol{\delta}}_k|\boldsymbol{\theta})}{p\left(\hat{\boldsymbol{\delta}}_k|\boldsymbol{\theta}, \mathbf{o}_k\right)} = exp(-1/2\, \hat{\boldsymbol{\delta}}_k \mathbf{K}^{-1}_{\boldsymbol{\delta},k} \hat{\boldsymbol{\delta}}_k) \approx exp(-1/2\, \hat{\boldsymbol{\delta}}_{s,k} \mathbf{K}^{-1}_{ss,k} \hat{\boldsymbol{\delta}}_{s,k}), \tag{A.15}$$

Basic synthetic example
clarifies effects of GP and imbalanced streams

Soil respiration measurements

- Rich: CO_2 production rate was measured over 7 days (~*1000* minutes) together with soil temperature
- Sparse: Each of *10* years, one monthly cumulated CO_2 was measured together with soil carbon

Two regressions with same coefficients, but with different covariates

Mean temperature

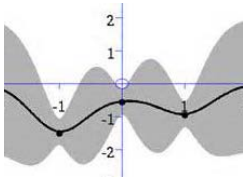
$$y_1 = a\, x_{\mathrm{Sparse}} + b\, \bar{x}_{\mathrm{Rich}}/10$$
$$y_2 = a\, x_{\mathrm{Sparse},1} + b\, (x_{\mathrm{Rich}} - \mathrm{c})$$

Single year soil carbon

Bayesian model inversion (MCMC) of a biased model
(fixed c = 0.2 instead of c = 0.3)

What you have heard

Gaussian Processes (GP) for discrepancy

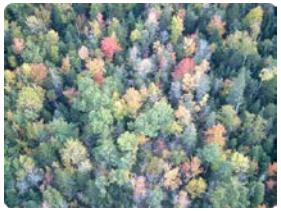


Basic example

- Increased uncertainty estimate
- Correctly located model discrepancy

Innovations for large data streams

- Discrepancy as hidden variable



Proof of concept on real world example