Modeling and simulation of high frequency wind vectors

Michael L. Stein

Rutgers University

Texas A&M, April 2022

ARM data

As part of DOE's Atmospheric Radiation Measurement program, surface meteorology has been measured every minute at Lamont, OK since 1993.

- Surface wind speed, wind direction, air temperature, relative humidity, barometric pressure, and precipitation
- Consistent instrumentation
- Low fraction missing

For several years, this information was collected at a network of sites in Southern Great Plains, but now only at this central facility

In 2001, I became head of the EPA-funded Center for Integrating Statistical and Environmental Science

- Initially had the ambition to analyze all the sites and all of the surface meteorology other than precipitation
- My main tool back then was Gaussian processes
- After looking at space-time patterns of winds, decided this was hopeless

Fast forward to 2021

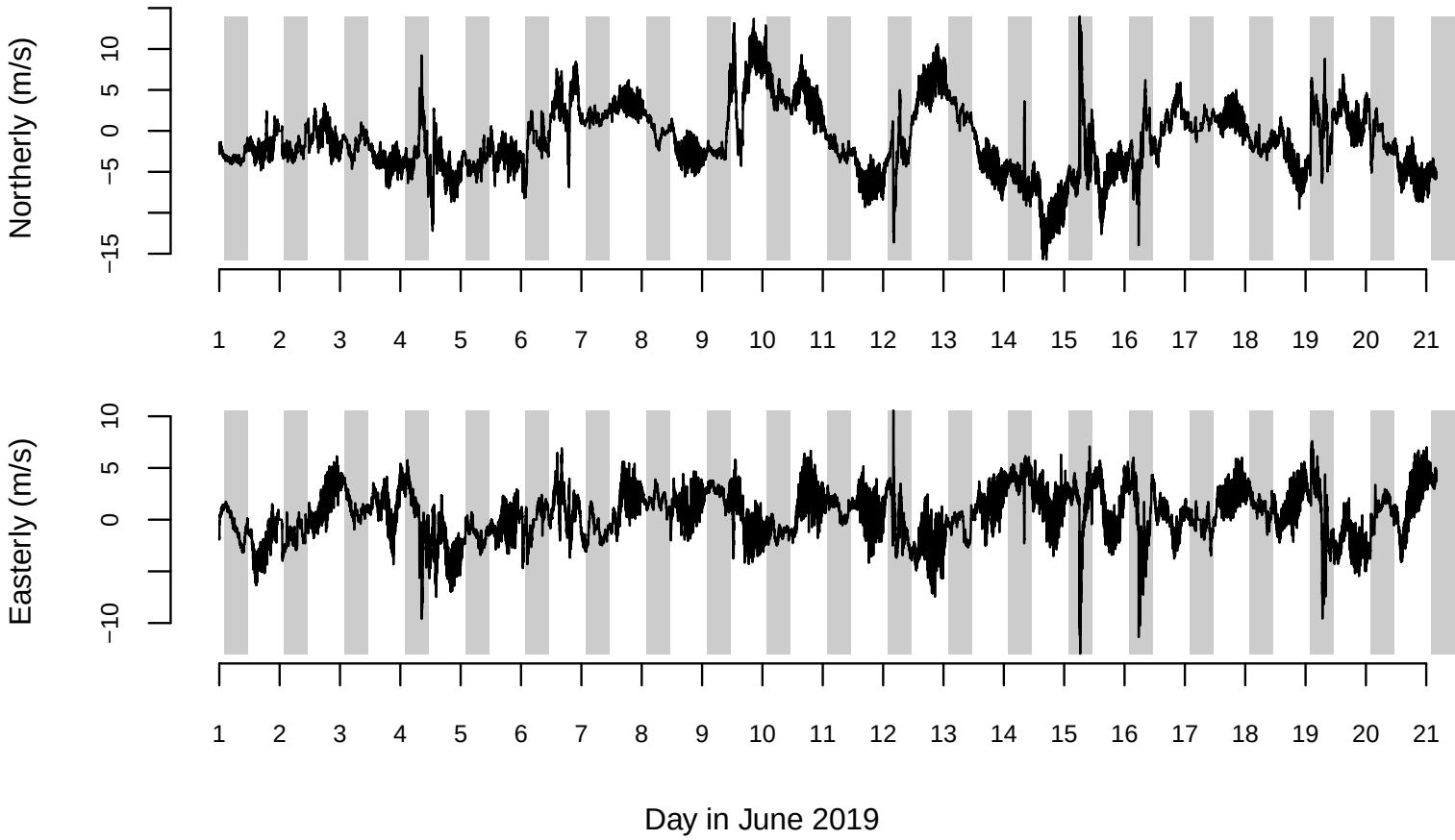
In recent years, my research has largely been on extremes

Main focus on extremes of a single quantity, but I wanted to move to multivariate extremes

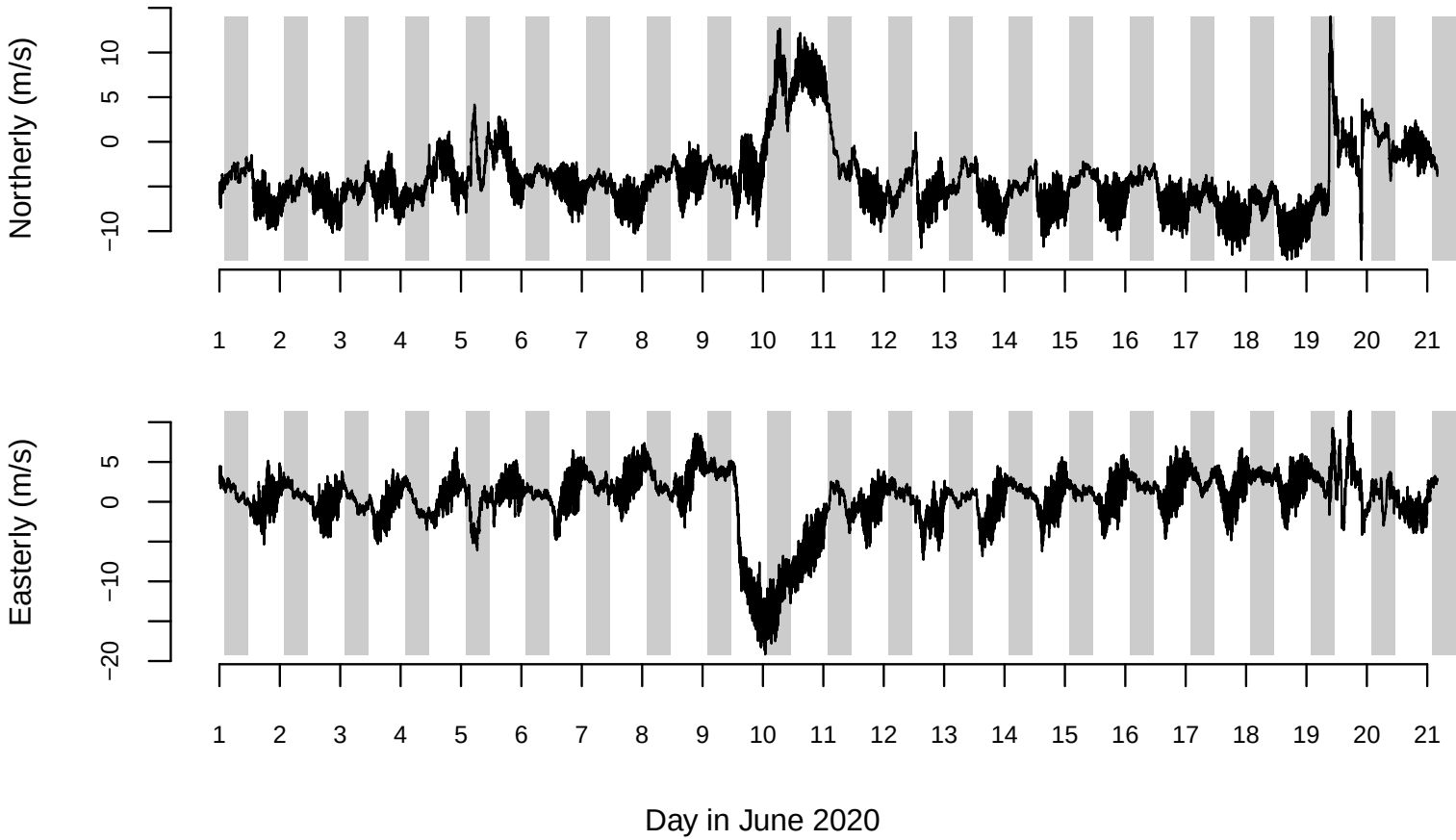
- I was dissatisfied with existing approaches for extremes dependence
- I thought focusing on a case where the vector was a physical vector (has a magnitude and a direction) could help in generating new ideas
- So I thought I would look at the ARM horizontal wind vector data

Maybe 20 years later and looking at only one rather than multiple sites, I could make some sense of it?

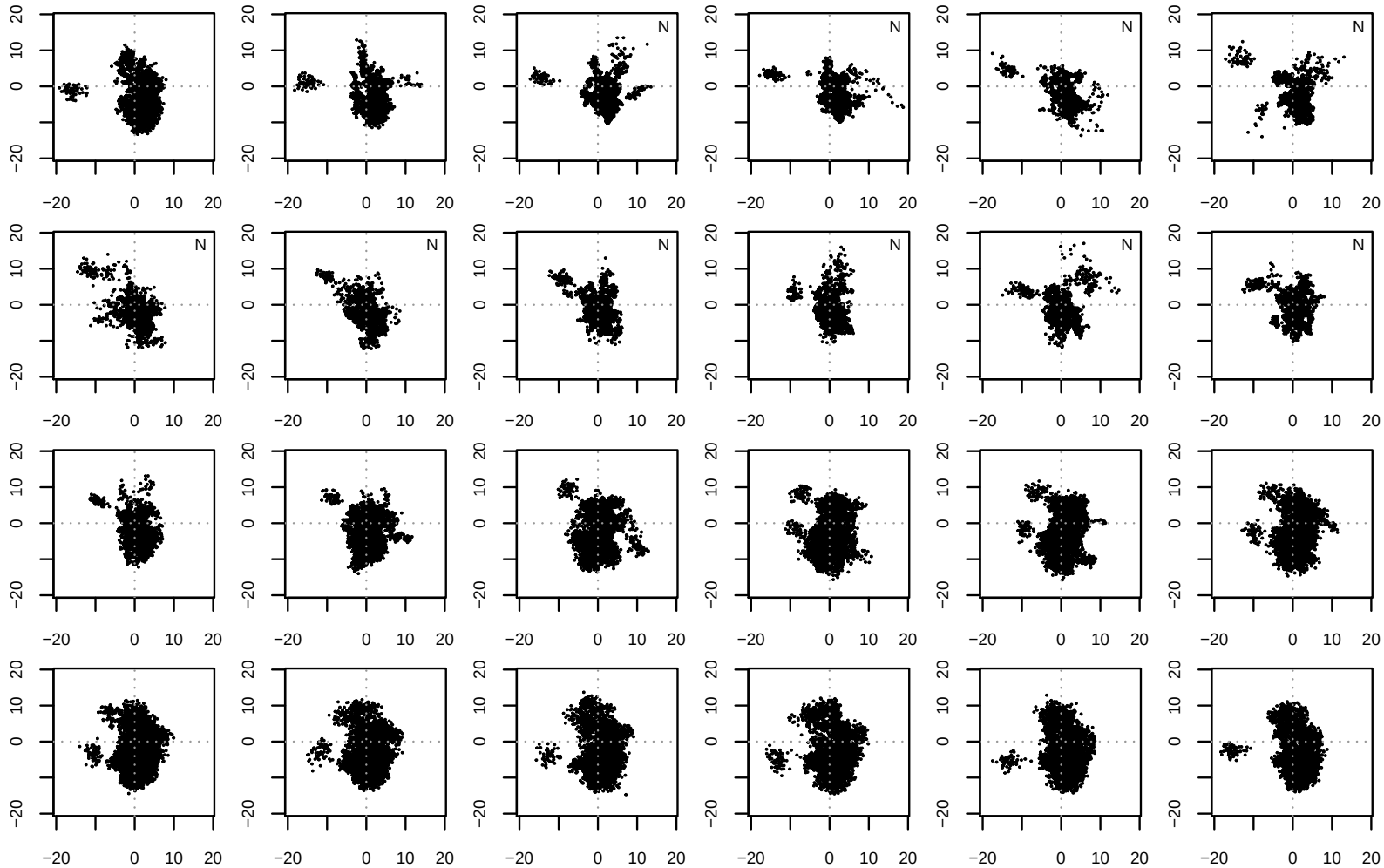
- Use vector means within each minute for wind speed and direction then convert to northerly and easterly components
- The data show tricky patterns depending strongly on whether it is night or day, so there is strong seasonal$\times$diurnal interaction
- To simplify, look at midnight June 1 through 4 am June 21 from 2016–2020 (times are UTC)
  - No missing observations helps
  - 145,200 observations, so need to use methods that can handle this much data



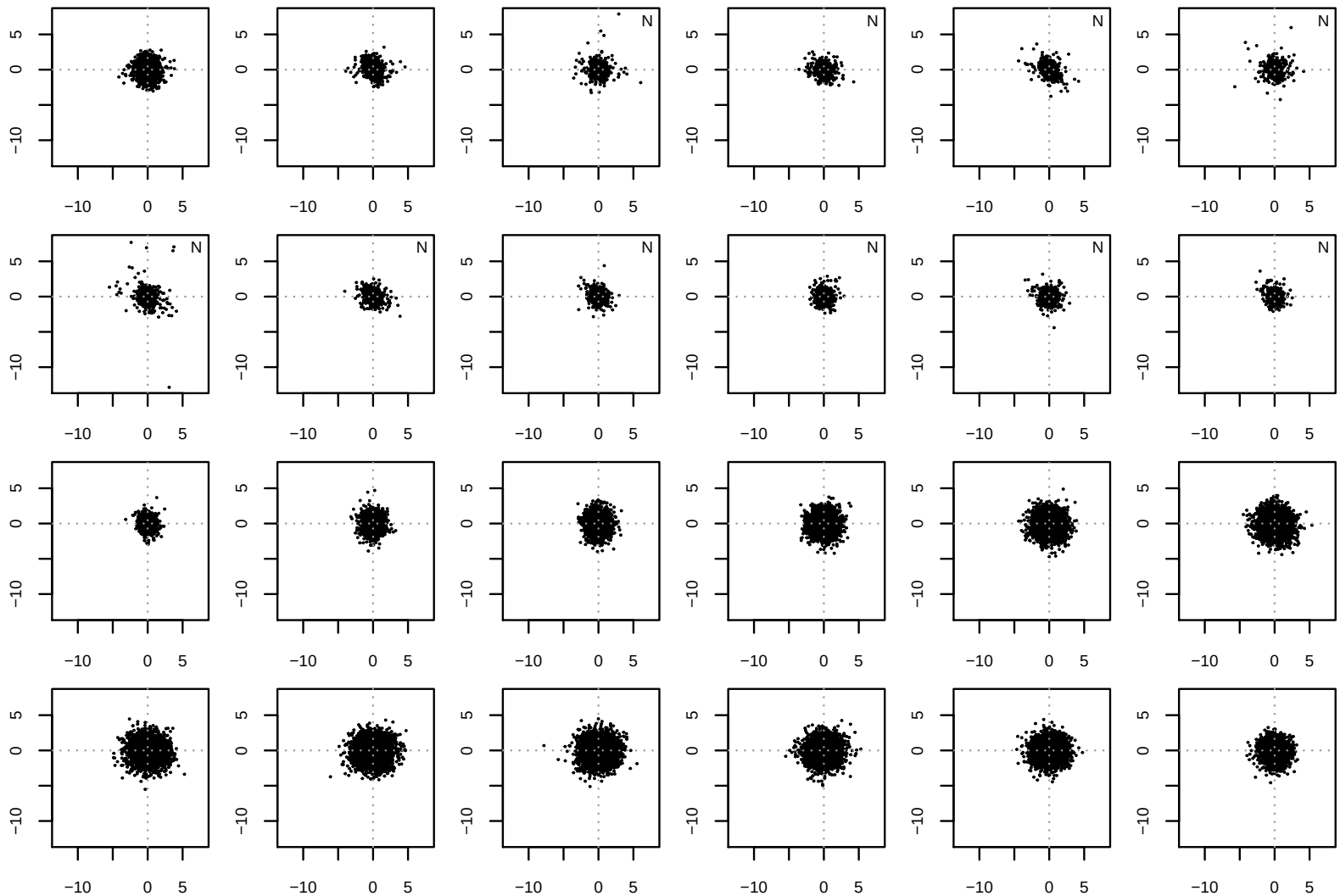
Wind vectors, 2019. Gray = nighttime.



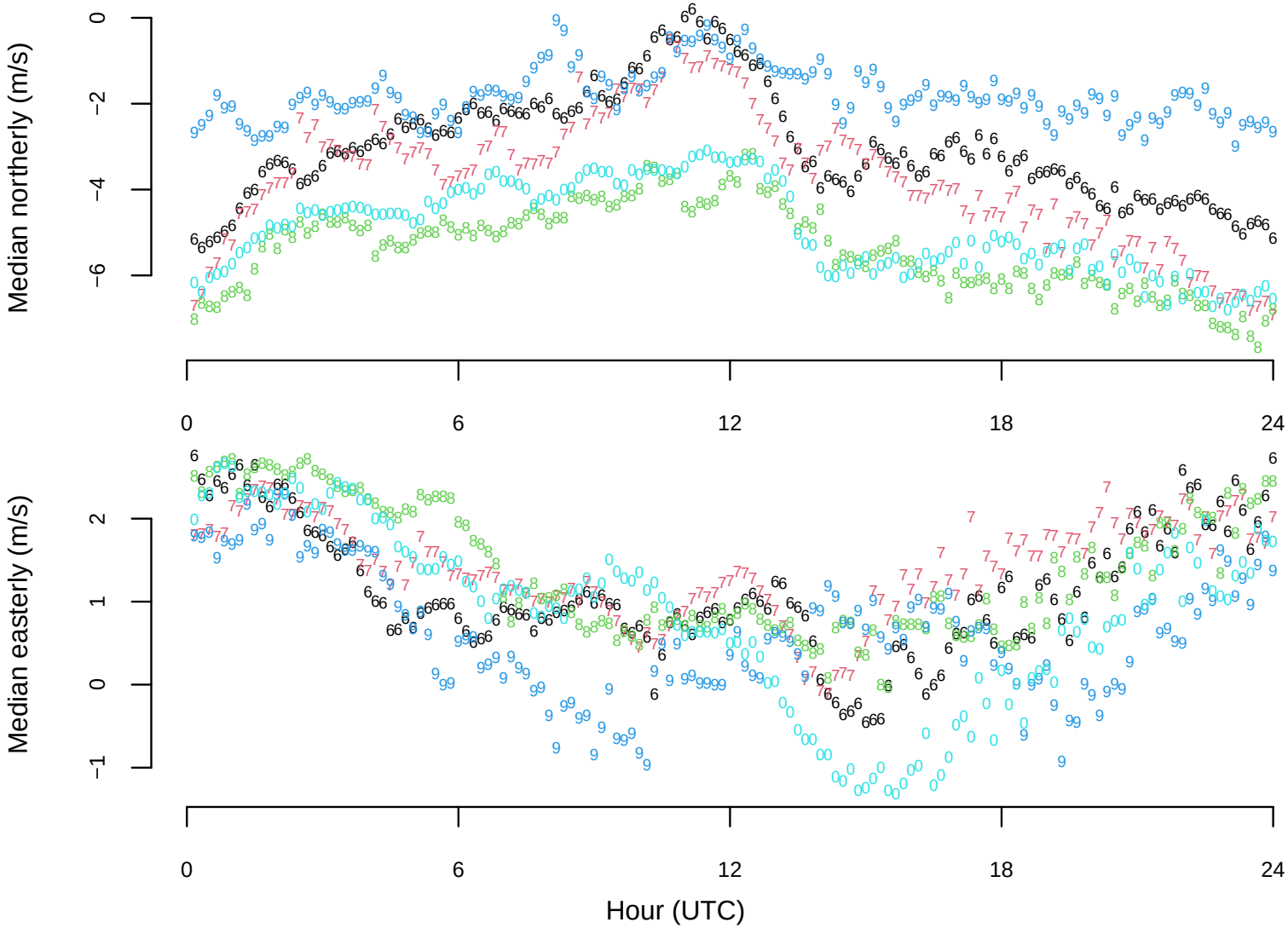
Wind vectors, 2020. Gray = nighttime.



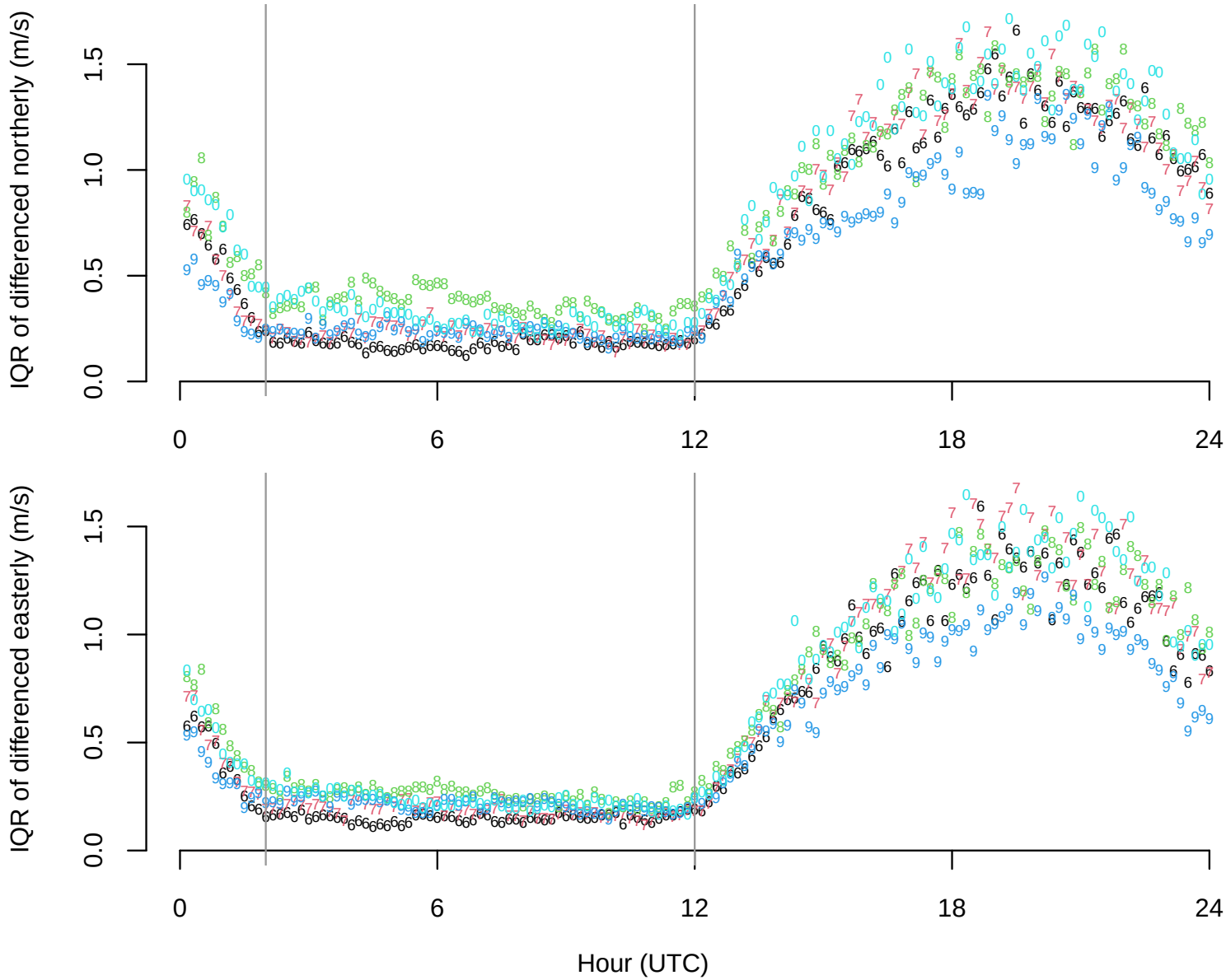
Wind vector (m/s) by hour (UTC). N indicates nighttime.



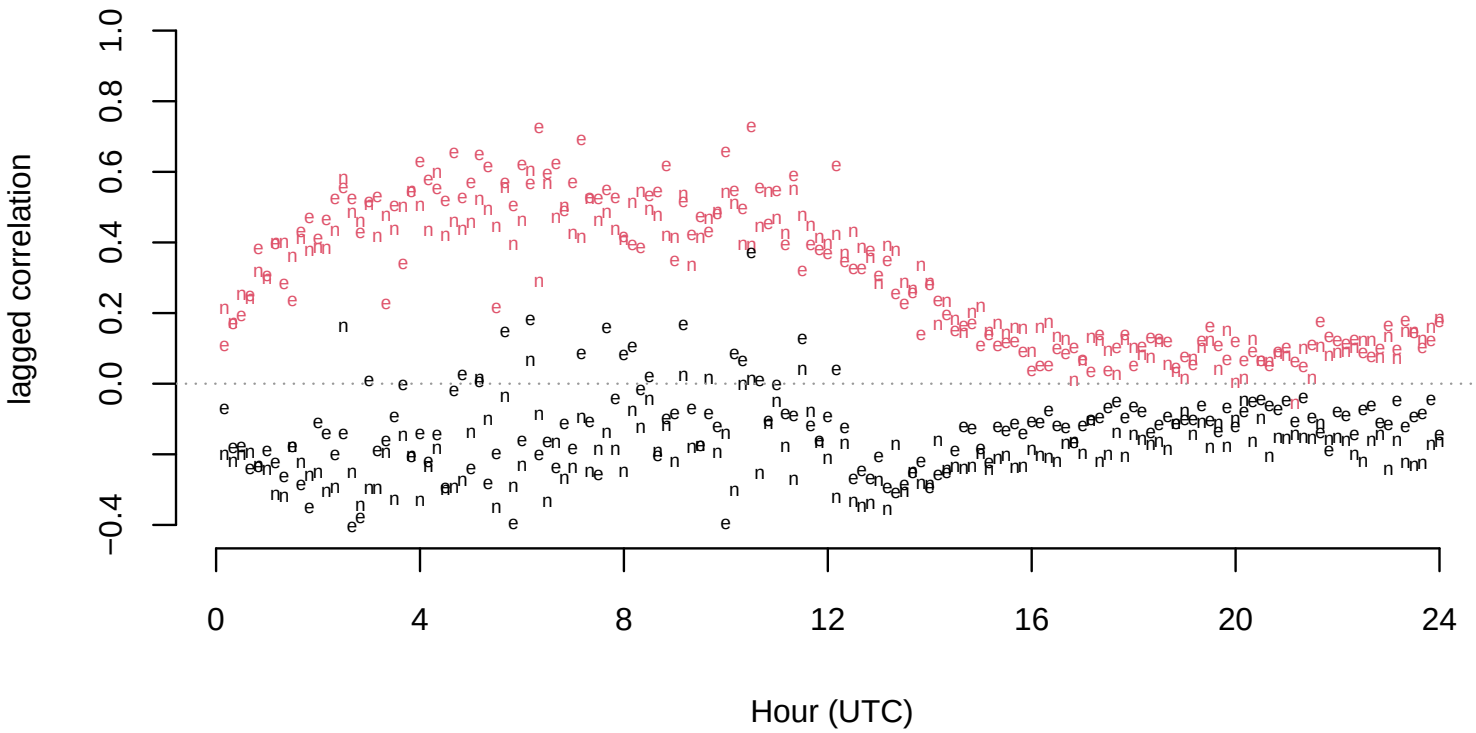
First differences in wind vector (m/s) by hour. N indicates nighttime.



Top: Medians of one-minute northerly wind for every 10-minute period. Plotting symbol = last digit of the year. Bottom: Same for easterly component.



Interquartile range of first differences. Gray lines delineate nighttime.



Correlations in first differences of wind vector (black) and absolute values of first differences (red). n = northerly component and e = easterly.

- Strong correlation in absolute differences sign of stochastic volatility: magnitude of changes in wind vector related to magnitude of past changes

Extremes

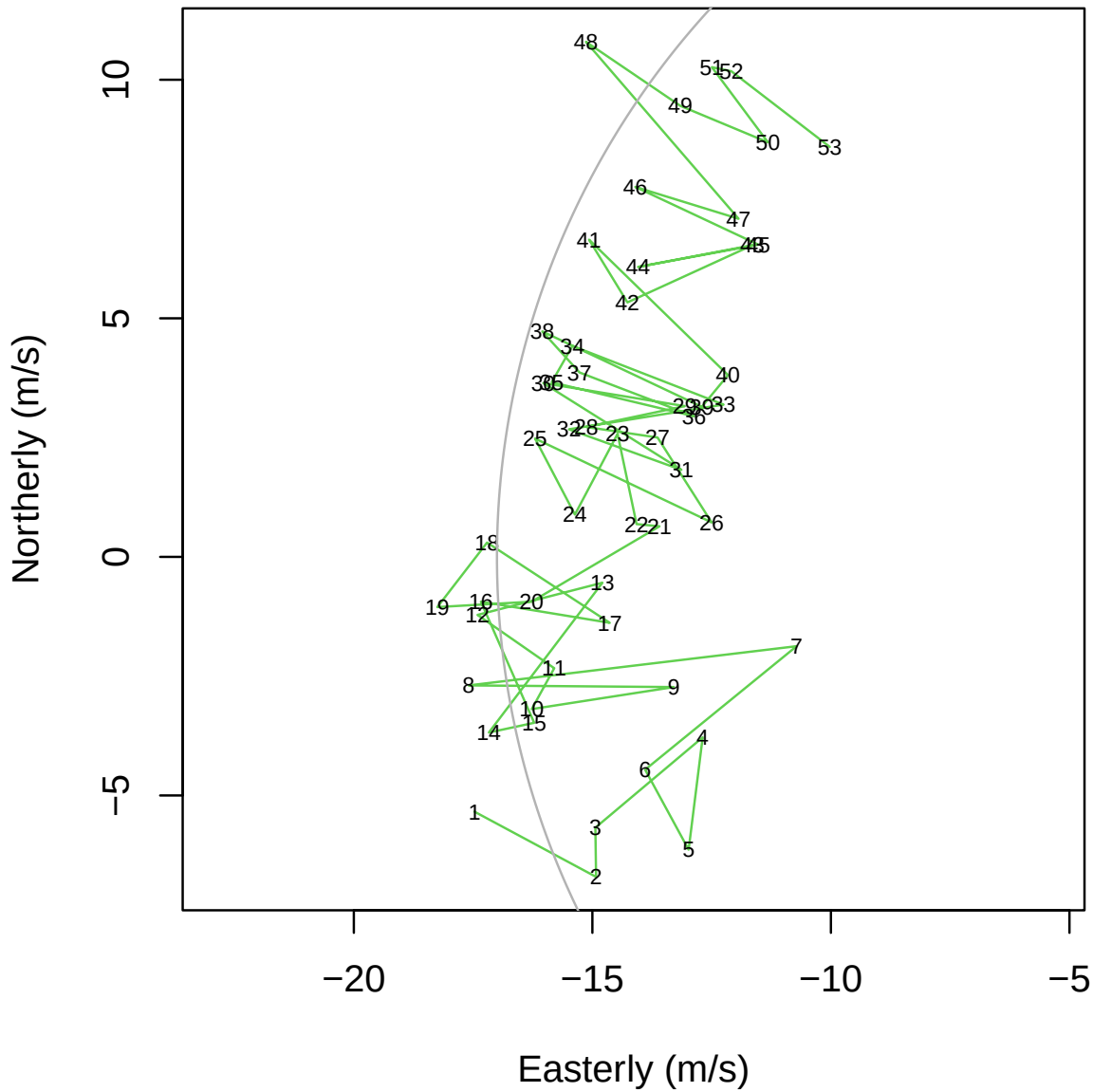
Some records in these data:

- Highest recorded wind speed: 19.75 m/s (44.2 mph)
- Highest recorded change in wind speed in one minute: 7.47 m/s
- Highest recorded change in wind vector in one minute: 13.22 m/s
- Lowest recorded wind speed: 0 m/s (640 times in 81 streaks from 1 to 89 minutes)
- Of 73 wind speeds > 17 m/s, number occurring during June 2–3, 2020: 66

These data cover too narrow a range of time to be of much use for inference on extremes

- At a site in Wichita, KS (about 70 miles away), record June wind from 1970–2018 is 83 mph (37 m/s)
- Of course, a tornado (and Lamont is right in the middle of Tornado Alley) can have much higher winds

Thus, while this dataset has lots of information for some aspects of process, it has limited information for others



Wind vector (every 10 minutes) during 9 hours of extreme event. Left of gray curve are winds $>$ 17 m/s.

Goal

Find a statistical model for the time series that yields realistic simulations

- Stochastic weather generators are popular as inputs into engineering models for weather impacts

When is wind vector, and not just wind speed, of interest?

- Wind energy? Maybe a bit
- Aviation. Wind direction relevant for takeoff and landing of planes
- Air pollution dispersal
- Perhaps would provide some useful insights for meteorologists?

Can any existing time series models capture all of the features previously noted?

Model the conditional distribution for the bivariate time series given the past of the series

- Turns out to be very challenging
- Modestly misspecifying conditional behavior can yield terrible simulations
- Simultaneously fitting all parts of model a difficult optimization problem

To allow broad exploration, use models that can be fairly easily fit

Model outline

I will use a three-stage modeling procedure:

1. Conditional median of each wind component
2. Conditional spread of each wind component

Use quantile regressions for first two stages to get normalized residuals with median 0 and constant spread

Many time series models assume these normalized innovations are independent and identically distributed. Badly untrue here.

Some possibilities for stage 3:

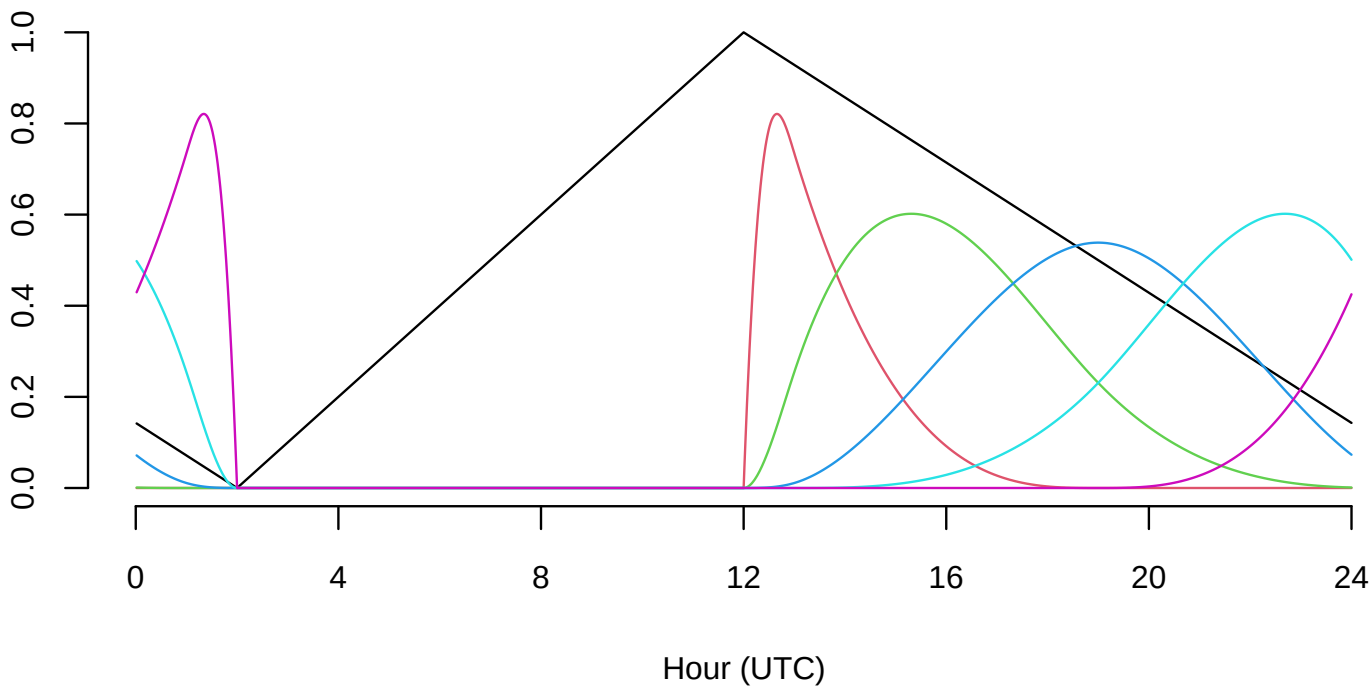
- Initial idea: Bivariate t with degrees of freedom and dependence parameter each have diurnal cycles. Doesn't suffice
- Empirical using estimated innovations depending on time of day and recent state of system
- Some smoothed version of this empirical distribution. I couldn't find any clear advantage over not smoothing

I will first describe the model, then try to explain where it came from

Conditional median

Notation:

- $V(t) = (N(t), E(t))$ for northerly and easterly components of wind at time t
- $\mathcal{F}(s)$ for wind data through time s
- m_t is the minute of day associated with time t, takes values $1, 2, \ldots, 1440$
- To capture (all) diurnal cycles, use basis functions a_1, constant, and $a_2, \ldots, a_7$ plotted here:



The conditional median for $N(t)$, denoted by $q_{0.5}(N(t) \mid \mathcal{F}(t-1))$, has the form

$$
\begin{aligned}
& q_{0.5}(N(t) \mid \mathcal{F}(t-1)) \\
&= \sum_{k=1}^{7} \theta_k a_k(m_t) + \sum_{k=1}^{7} \theta_{k+7} a_k(m_t) N(t-1) + \sum_{k=1}^{7} \theta_{k+14} a_k(m_t) N(t-2) \\
&\quad + \sum_{k=1}^{7} \theta_{k+21} a_k(m_t) \sum_{j=3}^{120} \delta^{j-3} N(t-j) \\
&\quad + \theta_{29} E(t-1) + \theta_{30} \tau \tan^{-1}\left(\frac{N(t-1)}{\tau}\right) + \theta_{31} \tau \tan^{-1}\left(\frac{N(t-2)}{\tau}\right) \\
&\quad + \theta_{32} \sum_{j=3}^{120} \delta^{j-3} \tau \tan^{-1}\left(\frac{N(t-j)}{\tau}\right)
\end{aligned}
$$

This model has a total of 34 parameters, the linear parameters $\theta_1, \ldots, \theta_{32}$ and the two nonlinear parameters $\delta \in (0,1)$ and $\tau > 0$

Similar model for $q_{0.5}(E(t) \mid \mathcal{F}(t-1))$

Interpretation of terms

$$
\begin{aligned}
&q_{0.5}(N(t) \mid \mathcal{F}(t-1)) \\
&= \underline{\sum_{k=1}^{7} \theta_k a_k(m_t)} + \sum_{k=1}^{7} \theta_{k+7} a_k(m_t) N(t-1) + \sum_{k=1}^{7} \theta_{k+14} a_k(m_t) N(t-2) \\
&\quad + \sum_{k=1}^{7} \theta_{k+21} a_k(m_t) \sum_{j=3}^{120} \delta^{j-3} N(t-j) \\
&\quad + \theta_{29} E(t-1) + \theta_{30} \tau \tan^{-1}\left(\frac{N(t-1)}{\tau}\right) + \theta_{31} \tau \tan^{-1}\left(\frac{N(t-2)}{\tau}\right) \\
&\quad + \theta_{32} \sum_{j=3}^{120} \delta^{j-3} \tau \tan^{-1}\left(\frac{N(t-j)}{\tau}\right)
\end{aligned}
$$

Controls unconditional median but is *not* even approximately the median

$$
\begin{aligned}
& q_{0.5}(N(t) \mid \mathcal{F}(t-1)) \\
&= \sum_{k=1}^{7} \theta_k a_k(m_t) + \underline{\sum_{k=1}^{7} \theta_{k+7} a_k(m_t) N(t-1) + \sum_{k=1}^{7} \theta_{k+14} a_k(m_t) N(t-2)} \\
&\quad + \sum_{k=1}^{7} \theta_{k+21} a_k(m_t) \sum_{j=3}^{120} \delta^{j-3} N(t-j) \\
&\quad + \theta_{29} E(t-1) + \theta_{30} \tau \tan^{-1}\left(\frac{N(t-1)}{\tau}\right) + \theta_{31} \tau \tan^{-1}\left(\frac{N(t-2)}{\tau}\right) \\
&\quad + \theta_{32} \sum_{j=3}^{120} \delta^{j-3} \tau \tan^{-1}\left(\frac{N(t-j)}{\tau}\right)
\end{aligned}
$$

AR(1) and AR(2) terms with diurnal cycles

$$
\begin{aligned}
& q_{0.5}(N(t) \mid \mathcal{F}(t-1)) \\
&= \sum_{k=1}^{7} \theta_k a_k(m_t) + \sum_{k=1}^{7} \theta_{k+7} a_k(m_t) N(t-1) + \sum_{k=1}^{7} \theta_{k+14} a_k(m_t) N(t-2) \\
&\quad + \underline{\sum_{k=1}^{7} \theta_{k+21} a_k(m_t) \sum_{j=3}^{120} \delta^{j-3} N(t-j)} \\
&\quad + \theta_{29} E(t-1) + \theta_{30} \tau \tan^{-1}\left(\frac{N(t-1)}{\tau}\right) + \theta_{31} \tau \tan^{-1}\left(\frac{N(t-2)}{\tau}\right) \\
&\quad + \theta_{32} \sum_{j=3}^{120} \delta^{j-3} \tau \tan^{-1}\left(\frac{N(t-j)}{\tau}\right)
\end{aligned}
$$

Cheap (computationally) approximation to MA term with diurnal cycle

$$
\begin{aligned}
&q_{0.5}(N(t) \mid \mathcal{F}(t-1)) \\
&= \sum_{k=1}^{7} \theta_k a_k(m_t) + \sum_{k=1}^{7} \theta_{k+7} a_k(m_t) N(t-1) + \sum_{k=1}^{7} \theta_{k+14} a_k(m_t) N(t-2) \\
&\quad + \sum_{k=1}^{7} \theta_{k+21} a_k(m_t) \sum_{j=3}^{120} \delta^{j-3} N(t-j) \\
&\quad + \underline{\theta_{29} E(t-1)} + \theta_{30} \tau \tan^{-1}\left(\frac{N(t-1)}{\tau}\right) + \theta_{31} \tau \tan^{-1}\left(\frac{N(t-2)}{\tau}\right) \\
&\quad + \theta_{32} \sum_{j=3}^{120} \delta^{j-3} \tau \tan^{-1}\left(\frac{N(t-j)}{\tau}\right)
\end{aligned}
$$

Impact of most recent value of other wind component

$$
\begin{aligned}
&q_{0.5}(N(t) \mid \mathcal{F}(t-1)) \\
&= \sum_{k=1}^{7} \theta_k a_k(m_t) + \sum_{k=1}^{7} \theta_{k+7} a_k(m_t) N(t-1) + \sum_{k=1}^{7} \theta_{k+14} a_k(m_t) N(t-2) \\
&\quad + \sum_{k=1}^{7} \theta_{k+21} a_k(m_t) \sum_{j=3}^{120} \delta^{j-3} N(t-j) \\
&\quad + \theta_{29} E(t-1) + \theta_{30} \tau \tan^{-1}\left(\frac{N(t-1)}{\tau}\right) + \theta_{31} \tau \tan^{-1}\left(\frac{N(t-2)}{\tau}\right) \\
&\quad + \theta_{32} \sum_{j=3}^{120} \delta^{j-3} \tau \tan^{-1}\left(\frac{N(t-j)}{\tau}\right)
\end{aligned}
$$

Nonlinear AR(1) and AR(2) terms with no diurnal cycle

- Nonlinear function is odd in $N(\cdot)$

$$
\begin{aligned}
&q_{0.5}(N(t) \mid \mathcal{F}(t-1)) \\
&= \sum_{k=1}^{7} \theta_k a_k(m_t) + \sum_{k=1}^{7} \theta_{k+7} a_k(m_t) N(t-1) + \sum_{k=1}^{7} \theta_{k+14} a_k(m_t) N(t-2) \\
&\quad + \sum_{k=1}^{7} \theta_{k+21} a_k(m_t) \sum_{j=3}^{120} \delta^{j-3} N(t-j) \\
&\quad + \theta_{29} E(t-1) + \theta_{30} \tau \tan^{-1}\left(\frac{N(t-1)}{\tau}\right) + \theta_{31} \tau \tan^{-1}\left(\frac{N(t-2)}{\tau}\right) \\
&\quad + \underline{\theta_{32} \sum_{j=3}^{120} \delta^{j-3} \tau \tan^{-1}\left(\frac{N(t-j)}{\tau}\right)}
\end{aligned}
$$

Nonlinear "MA" term with no diurnal cycle. Note same δ as for linear MA term.

Conditional spread

Define $\bar{N}(t) = N(t) - q_{0.5}(N(t) \mid \mathcal{F}(t-1))$ and $W(t) = |V(t)|$ is wind speed (m/s) at time t.

Model conditional spread through quantile of $|\bar{N}(t)|$:

$$
\begin{aligned}
& q_{0.9}(|\bar{N}(t)| \mid \mathcal{F}(t-1)) \\
&= \sum_{k=1}^{7} \phi_k a_k(m_t) \Bigg[\sum_{k=1}^{7} \phi_{k+7} a_k(m_t) |\bar{N}(t-1)| + \phi_{15} |\bar{N}(t-2)| + \phi_{16} |\bar{N}(t-3)| \\
&\qquad + \phi_{17} \sum_{j=4}^{120} \gamma_1^{j-4} |\bar{N}(t-j)| + \phi_{18} |\bar{E}(t-1)| \\
&\qquad + \phi_{19} \sum_{j=2}^{120} \gamma_2^{j-2} |\bar{E}(t-j)| + \sum_{k=1}^{7} \phi_{19+k} a_k(m_t) \frac{W(t-1)}{\sigma + W(t-1)} \Bigg]
\end{aligned}
$$

This model has a total of 29 parameters, $\phi_1, \ldots, \phi_{26}$ and the 3 nonlinear parameters $\gamma_1 \in (0,1)$, $\gamma_2 \in (0,1)$ and $\sigma > 0$.

- First estimate $\phi_1, \ldots, \phi_7$. Then estimate all the others so never fit more than 3 nonlinear parameters.

Similar model for $q_{0.9}(|\bar{E}(t)| \mid \mathcal{F}(t-1))$

Interpretation of terms

$$
\begin{aligned}
&q_{0.9}(|\bar{N}(t)| \mid \mathcal{F}(t-1)) \\
&= \underline{\sum_{k=1}^{7} \phi_k a_k(m_t)} \Bigg[\sum_{k=1}^{7} \phi_{k+7} a_k(m_t) |\bar{N}(t-1)| + \phi_{15} |\bar{N}(t-2)| + \phi_{16} |\bar{N}(t-3)| \\
&\qquad + \phi_{17} \sum_{j=4}^{120} \gamma_1^{j-4} |\bar{N}(t-j)| + \phi_{18} |\bar{E}(t-1)| \\
&\qquad + \phi_{19} \sum_{j=2}^{120} \gamma_2^{j-2} |\bar{E}(t-j)| + \sum_{k=1}^{7} \phi_{19+k} a_k(m_t) \frac{W(t-1)}{\sigma + W(t-1)} \Bigg]
\end{aligned}
$$

Non-random diurnal volatility. Define as $DV_N(m_t)$.

Everything not underlined is stochastic volatility, $SV_N(t)$.

- Define total volatility $TV_N(t) = DV_N(m_t) \times SV_N(t)$
- Essential that DV_N multiplies SV_N

Similarly define $DV_E(m_t)$, $SV_E(t)$, $TV_E(t)$.

$$q_{0.9}(|\bar{N}(t)| \mid \mathcal{F}(t-1))$$

$$= \sum_{k=1}^{7} \phi_k a_k(m_t) \Bigg[\sum_{k=1}^{7} \phi_{k+7} a_k(m_t) |\bar{N}(t-1)| + \phi_{15} |\bar{N}(t-2)| + \phi_{16} |\bar{N}(t-3)|$$

$$+ \phi_{17} \sum_{j=4}^{120} \gamma_1^{j-4} |\bar{N}(t-j)| + \phi_{18} |\bar{E}(t-1)|$$

$$+ \phi_{19} \sum_{j=2}^{120} \gamma_2^{j-2} |\bar{E}(t-j)| + \sum_{k=1}^{7} \phi_{19+k} a_k(m_t) \frac{W(t-1)}{\sigma + W(t-1)} \Bigg]$$

AR(1) term with diurnal cycle

$$
\begin{aligned}
&q_{0.9}(|\bar{N}(t)| \mid \mathcal{F}(t-1)) \\
&= \sum_{k=1}^{7} \phi_k a_k(m_t) \Bigg[\sum_{k=1}^{7} \phi_{k+7} a_k(m_t) |\bar{N}(t-1)| + \underline{\phi_{15}|\bar{N}(t-2)| + \phi_{16}|\bar{N}(t-3)|} \\
&\qquad + \phi_{17} \sum_{j=4}^{120} \gamma_1^{j-4} |\bar{N}(t-j)| + \phi_{18} |\bar{E}(t-1)| \\
&\qquad + \phi_{19} \sum_{j=2}^{120} \gamma_2^{j-2} |\bar{E}(t-j)| + \sum_{k=1}^{7} \phi_{19+k} a_k(m_t) \frac{W(t-1)}{\sigma + W(t-1)} \Bigg]
\end{aligned}
$$

AR(2) and AR(3) terms without diurnal cycles

$$
\begin{aligned}
&q_{0.9}(|\bar{N}(t)| \mid \mathcal{F}(t-1)) \\
&= \sum_{k=1}^{7} \phi_k a_k(m_t) \Bigg[\sum_{k=1}^{7} \phi_{k+7} a_k(m_t) |\bar{N}(t-1)| + \phi_{15} |\bar{N}(t-2)| + \phi_{16} |\bar{N}(t-3)| \\
&\qquad + \underline{\phi_{17} \sum_{j=4}^{120} \gamma_1^{j-4} |\bar{N}(t-j)|} + \phi_{18} |\bar{E}(t-1)| \\
&\qquad + \phi_{19} \sum_{j=2}^{120} \gamma_2^{j-2} |\bar{E}(t-j)| + \sum_{k=1}^{7} \phi_{19+k} a_k(m_t) \frac{W(t-1)}{\sigma + W(t-1)} \Bigg]
\end{aligned}
$$

MA-like term without diurnal cycle

$$
\begin{aligned}
&q_{0.9}(|\bar{N}(t)| \mid \mathcal{F}(t-1)) \\
&= \sum_{k=1}^{7} \phi_k a_k(m_t) \Bigg[\sum_{k=1}^{7} \phi_{k+7} a_k(m_t) |\bar{N}(t-1)| + \phi_{15} |\bar{N}(t-2)| + \phi_{16} |\bar{N}(t-3)| \\
&\qquad + \phi_{17} \sum_{j=4}^{120} \gamma_1^{j-4} |\bar{N}(t-j)| + \underline{\phi_{18} |\bar{E}(t-1)|} \\
&\qquad + \phi_{19} \sum_{j=2}^{120} \gamma_2^{j-2} |\bar{E}(t-j)| + \sum_{k=1}^{7} \phi_{19+k} a_k(m_t) \frac{W(t-1)}{\sigma + W(t-1)} \Bigg]
\end{aligned}
$$

Effect of most recent absolute value of other centered wind component

$$
\begin{aligned}
&q_{0.9}(|\bar{N}(t)| \mid \mathcal{F}(t-1)) \\
&= \sum_{k=1}^{7} \phi_k a_k(m_t) \Bigg[\sum_{k=1}^{7} \phi_{k+7} a_k(m_t) |\bar{N}(t-1)| + \phi_{15} |\bar{N}(t-2)| + \phi_{16} |\bar{N}(t-3)| \\
&\qquad + \phi_{17} \sum_{j=4}^{120} \gamma_1^{j-4} |\bar{N}(t-j)| + \phi_{18} |\bar{E}(t-1)| \\
&\qquad + \underline{\phi_{19} \sum_{j=2}^{120} \gamma_2^{j-2} |\bar{E}(t-j)|} + \sum_{k=1}^{7} \phi_{19+k} a_k(m_t) \frac{W(t-1)}{\sigma + W(t-1)} \Bigg]
\end{aligned}
$$

MA-like effect of absolute value of other centered wind component

$$
\begin{aligned}
&q_{0.9}(|\bar{N}(t)| \mid \mathcal{F}(t-1)) \\
&= \sum_{k=1}^{7} \phi_k a_k(m_t) \Bigg[\sum_{k=1}^{7} \phi_{k+7} a_k(m_t) |\bar{N}(t-1)| + \phi_{15} |\bar{N}(t-2)| + \phi_{16} |\bar{N}(t-3)| \\
&\qquad + \phi_{17} \sum_{j=4}^{120} \gamma_1^{j-4} |\bar{N}(t-j)| + \phi_{18} |\bar{E}(t-1)| \\
&\qquad + \phi_{19} \sum_{j=2}^{120} \gamma_2^{j-2} |\bar{E}(t-j)| + \sum_{k=1}^{7} \phi_{19+k} a_k(m_t) \frac{W(t-1)}{\sigma + W(t-1)} \Bigg]
\end{aligned}
$$

Nonlinear effect of most recent wind speed with diurnal cycle

- In part to approximate behavior when $W(t-1) = 0$, but also lower volatility when $W(t-1)$ is low

Innovation distribution

Because of dependence in the two components of innovations at night, must handle them simultaneously

Define the empirical innovation

$$I_o(t) = (\tilde{N}(t), \tilde{E}(t))$$

where, e.g.,

$$\tilde{N}(t) = \bar{N}(t)/q_{0.9}(|\bar{N}(t)| \mid \mathcal{F}(t-1))$$

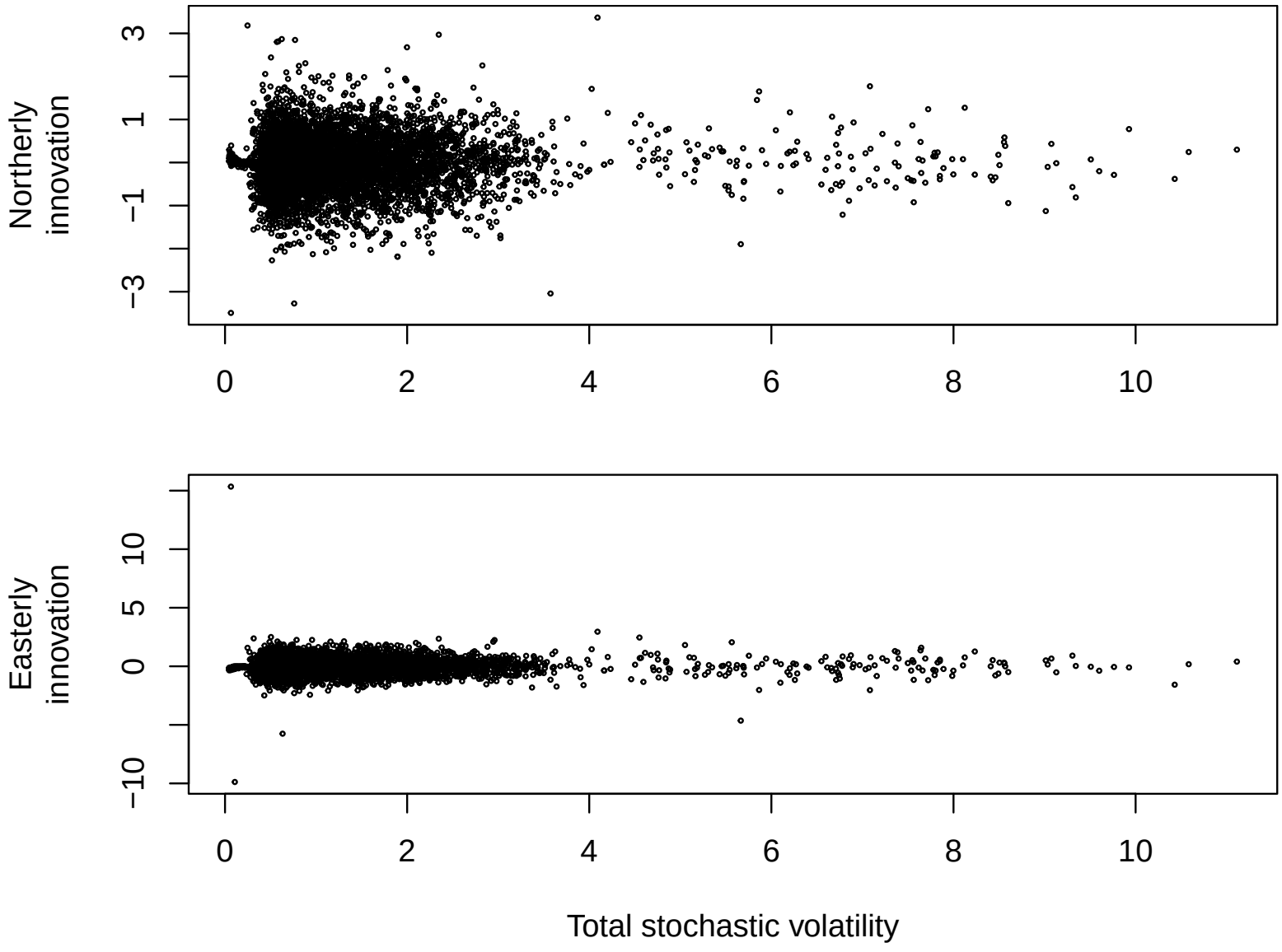
- Use subscript o to emphasize quantities defined in terms of observations
- Quantities without subscript o might be observations or simulations

Also need residuals only normalized for unconditional spread,

$$J_o(t) = (\bar{N}(t)/DV_N(m_t), \bar{E}(t)/DV_E(m_t))$$

Had hoped could approximate sampling distribution of innovations at time t by the empirical distribution of $I_o(s)$ for $|s - t| \leq 30 \pmod{1440}$, say

However, even these normalized innovations have a distribution depending on more than just the time of day



Innovations v $SV(t) = \sqrt{SV_N(t)^2 + SV_E(t)^2}$ for 6-7 UTC

Procedure for sampling innovations

Define

- $DV(m_t) = \sqrt{DV_N(m_t)^2 + DV_E(m_t)^2}$
- $TV(t) = \sqrt{TV_N(t)^2 + TV_E(t)^2}$

First separate out cases with high potential for extreme windspeed:

- Select cutoffs $\kappa_1 > \kappa_2 > \kappa_3$
- Compute $P(t) = q_{0.5}(N(t) \mid \mathcal{F}(t-1))^2 + q_{0.5}(E(t) \mid \mathcal{F}(t-1))^2 + TV(t)^2$
- If $P(t) > \kappa_1$, sample from $J_o(s)$ for which $P_o(s) > \kappa_1$
- Else if $P(t) > \kappa_2$, sample from $J_o(s)$ for which $\kappa_1 \geq P_o(s) > \kappa_2$
- Else if $P(t) > \kappa_3$, sample from $J_o(s)$ for which $\kappa_2 \geq P_o(s) > \kappa_3$

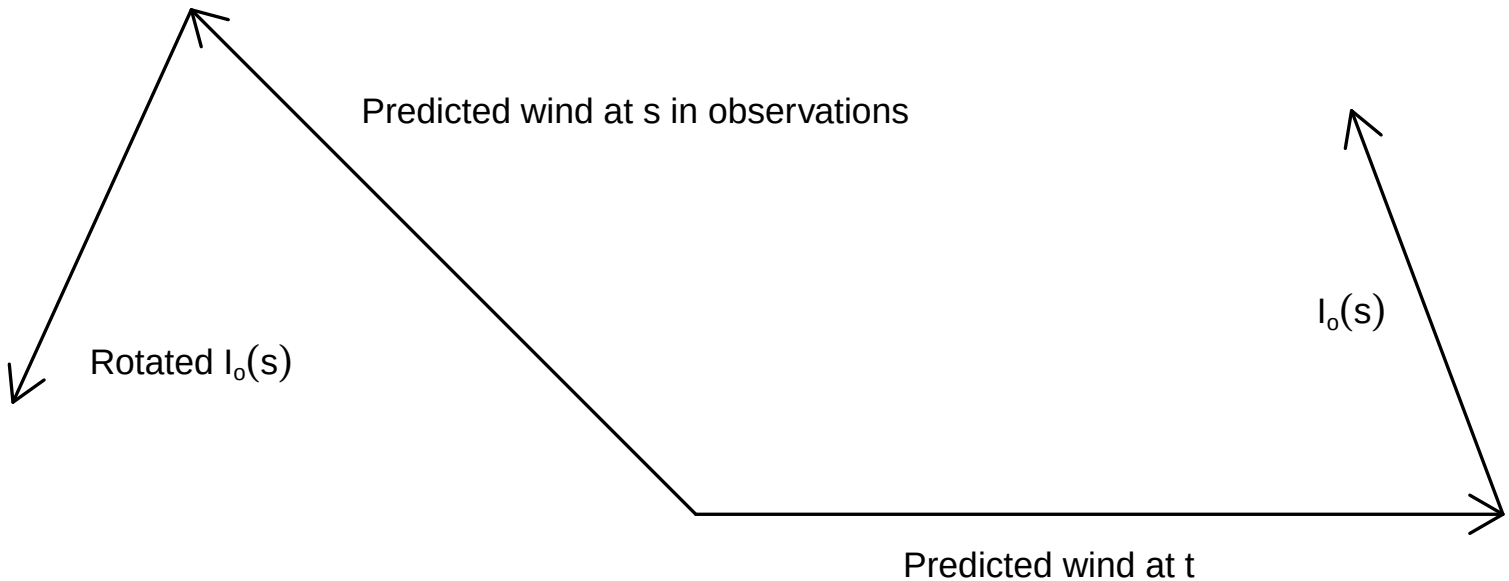
Next, do separate sampling from $I_o(s)$ for times of especially low ($< \lambda_1$) or high ($> \lambda_2$) stochastic volatility

- Else If $SV(t) < \lambda_1$, sample from $I_o(s)$ for which $SV_o(s) < \lambda_1$ and $P_o(s) \leq \kappa_3$.
- Else If $SV(t) > \lambda_2$, sample from $I_o(s)$ for which $SV_o(s) > \lambda_2$ and $P_o(s) \leq \kappa_3$.

For remaining cases, sample from $I_o(s)$ for which

- $P_o(s)$ and $SV_o(s)$ not in previous categories
- $|s - t| \leq 20 \pmod{1440}$
- $SV_o(s)$ is in same quintile of stochastic volatilities for that time window as $SV(t)$

Then rotate $I_o(s)$ or $J_o(s)$:



Simulation procedure

Suppose have $V_{t-240}, \ldots, V_{t-1}$ and want to simulate V_t

- From $V_{t-240}, \ldots, V_{t-1}$, compute conditional medians for each wind component for times $t - 120, \ldots, t$
- Compute $SV_N(t - 120), \ldots, SV_N(t)$ and $SV_E(t - 120), \ldots, SV_E(t)$
- Sample a normalized innovation and rotate it
- Put back in volatility (either DV or TV)
- Add back in conditional medians

To approximate conditional distribution of $V(t)$ given the past, just sample many normalized innovations

Where did I get this?

I tried *many* models for conditional mean, conditional spread and conditional innovations

Criteria

- As much as possible, use linear quantile regression
  - Quantile regression rather than least squares because of fat tails, especially at night
  - `method = 'fn'` in `rq` command in `quantreg` package in `R` is very fast
  - Fit nonlinear parameters in outer loop of nested optimization
- Thus, limit number of nonlinear parameters
  - Mimic MA behavior by including exponential decay of AR parameters
  - Do not model unconditional median by subtracting from $N(t)$ (or $E(t)$)
  - Fit conditional spread model by first fitting DV terms, then fit SV terms
- In quantile regressions, favor including terms leading to larger reductions in criterion function
- (Unweighted) Continuous ranked probability scores on one-minute ahead predictions
- Simulations look "realistic"
  - Biggest challenge. First simulations produced wind speeds $>$ speed of light
  - Scheme for sampling and rotating innovations largely devised to avoid unreasonable wind speeds

Complexity and simplicity

Seek a model that is simple as possible, but complex where needed

- Complex: Basis functions for diurnal effects flexible and cognizant of nighttime v. daytime
- Simple: Same basis functions for all diurnal effects

- Complex: Nonlinear AR terms in conditional mean
- Simple: Same nonlinear function with one parameter, τ, for all lags

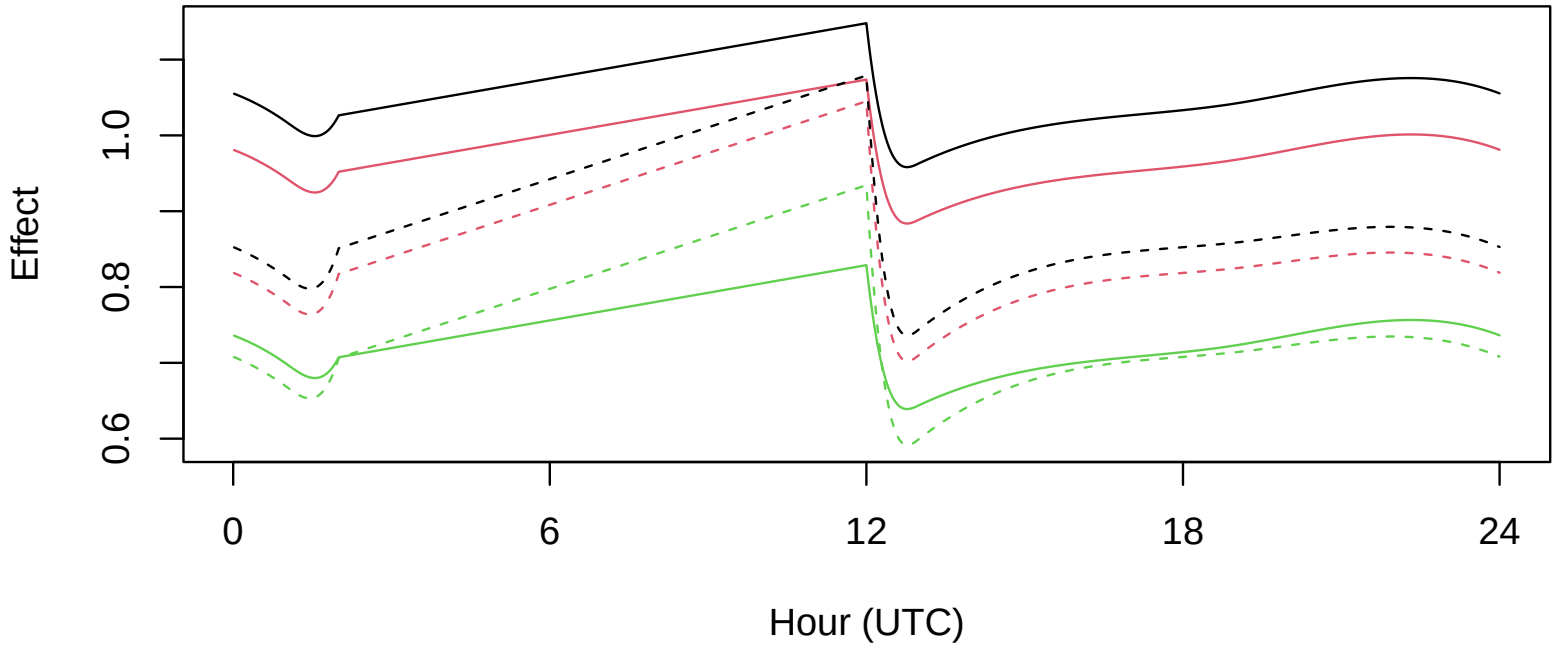
- Complex: Nonlinear function of $W(t-1)$ in conditional spread
- Simple: Nonlinear function with one parameter, σ

- Complex: Fairly elaborate forms for conditional median and spread
- Simple: Same form for northerly and easterly components

- Complex: Stratified scheme for sampling innovations
- Simple: Cutoffs chosen informally based on plots and quality of simulations

- Complex and simple: Rotating innovations

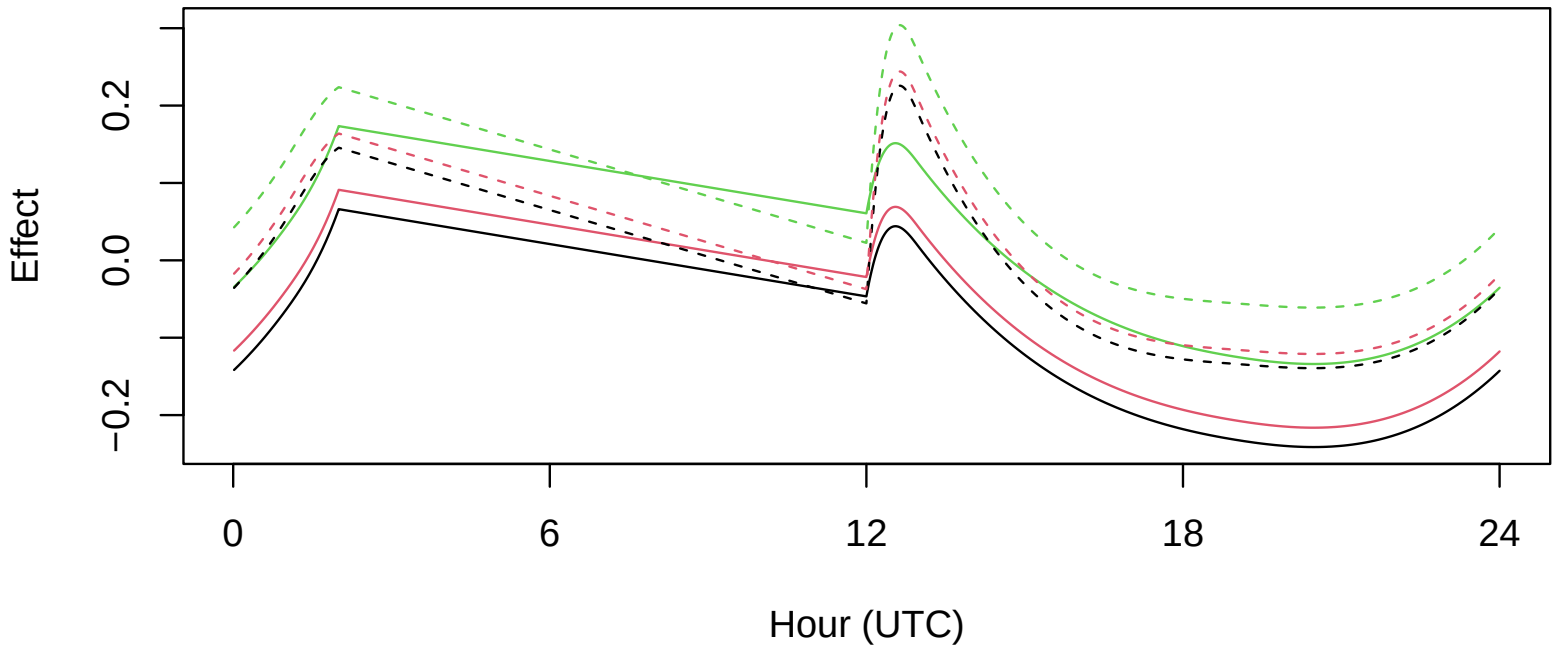
But the simplicity is *not* due to sparseness, at least not in ordinary sense.



Estimated AR(1) effects as function of time of day for northerly (solid) and easterly (dashed) winds

- Each curve shows the contribution to the one-minute ahead predicted median for three values of the wind vector:
  1 m/s (black), 3 m/s (red) and 10 m/s (green) scaled so that the curves would coincide if the effect were linear

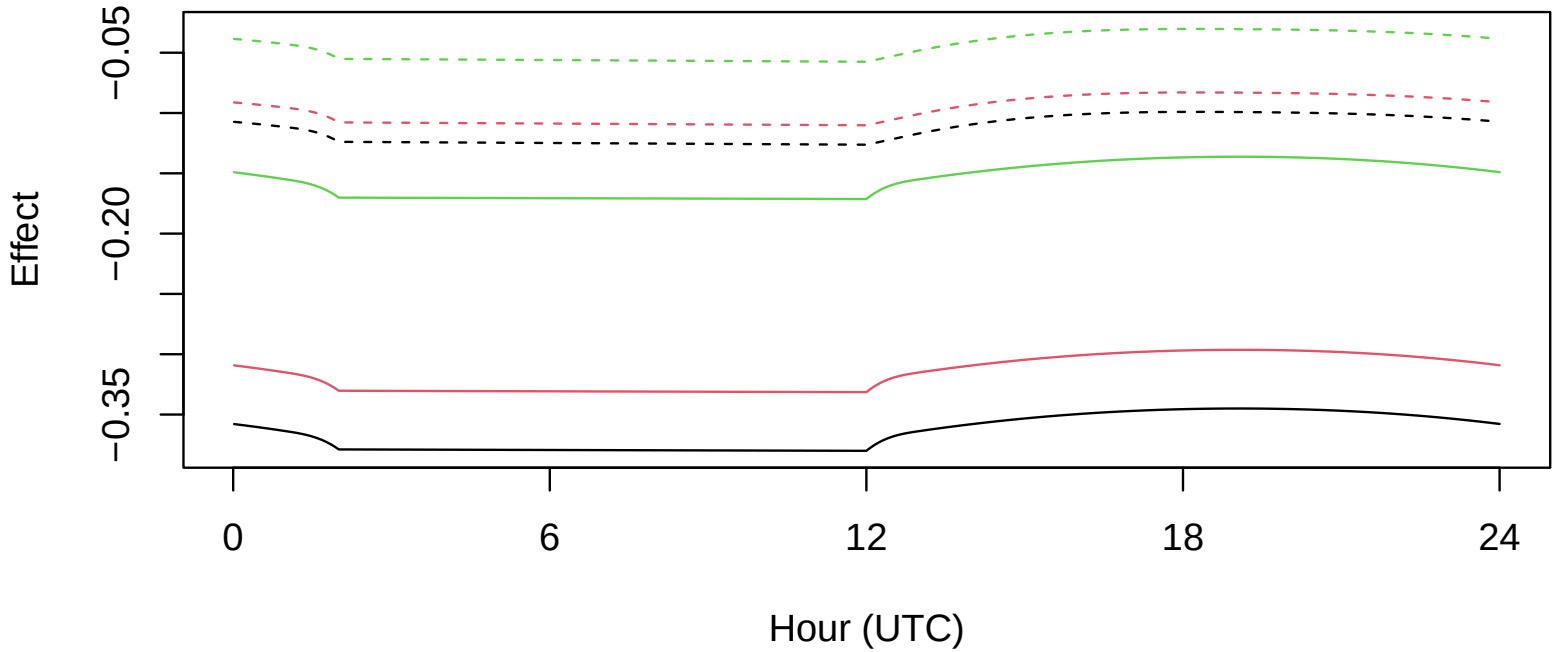
Nonlinearity weaker for easterly



Estimated AR(2) effects as function of time of day for northerly (solid) and easterly (dashed) winds

- Each curve shows the contribution to the one-minute ahead predicted median for three values of the wind vector:
  1 m/s (black), 3 m/s (red) and 10 m/s (green) scaled so that the curves would coincide if the effect were linear

Some aspects of fitted conditional median

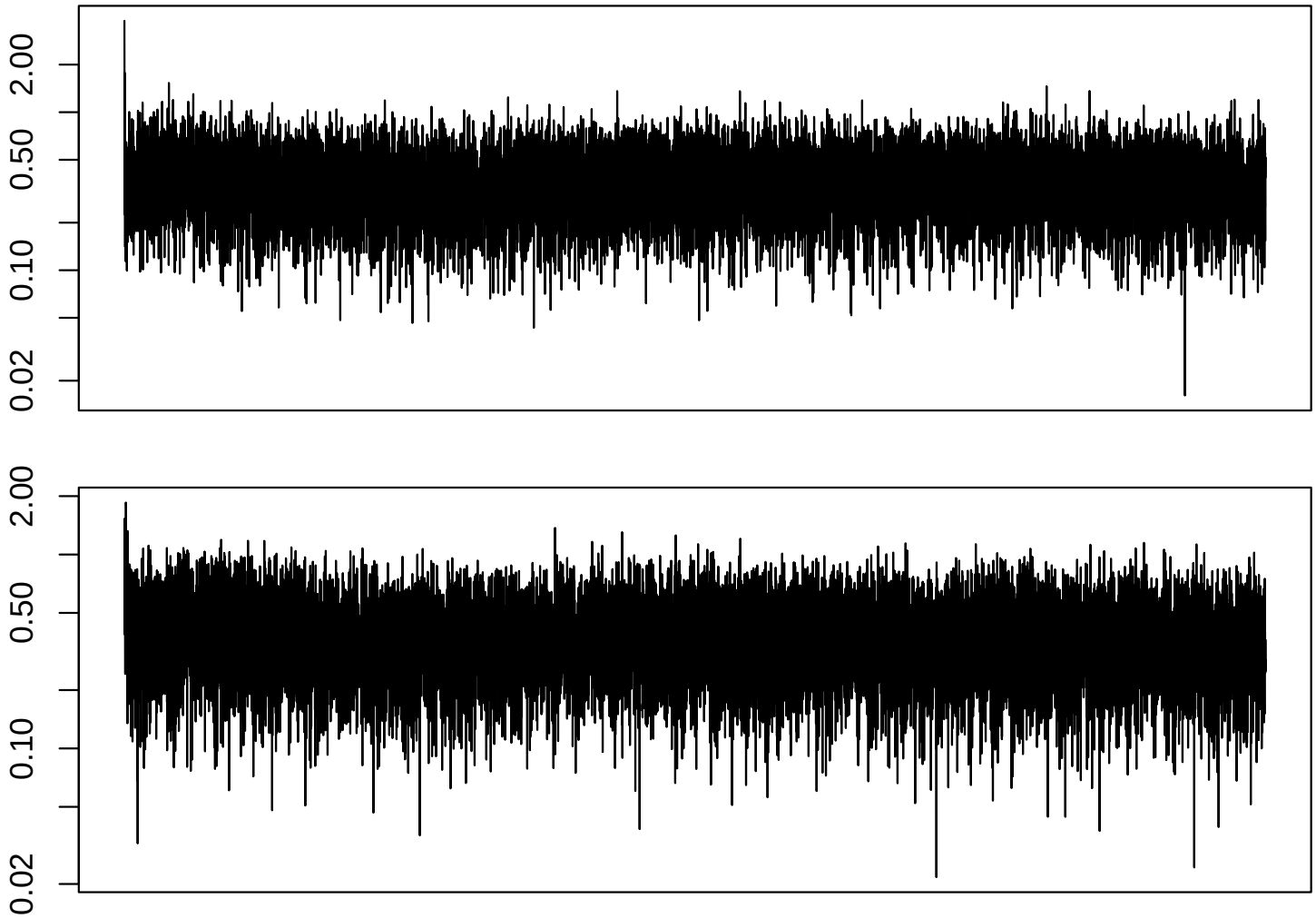


Estimated higher order effects as function of time of day for northerly (solid) and easterly (dashed) winds

- Each curve shows the contribution to the one-minute ahead predicted median for three values of the wind vector:
  1 m/s (black), 3 m/s (red) and 10 m/s (green) for lags 3–120 scaled so that the curves would coincide if the effect were linear

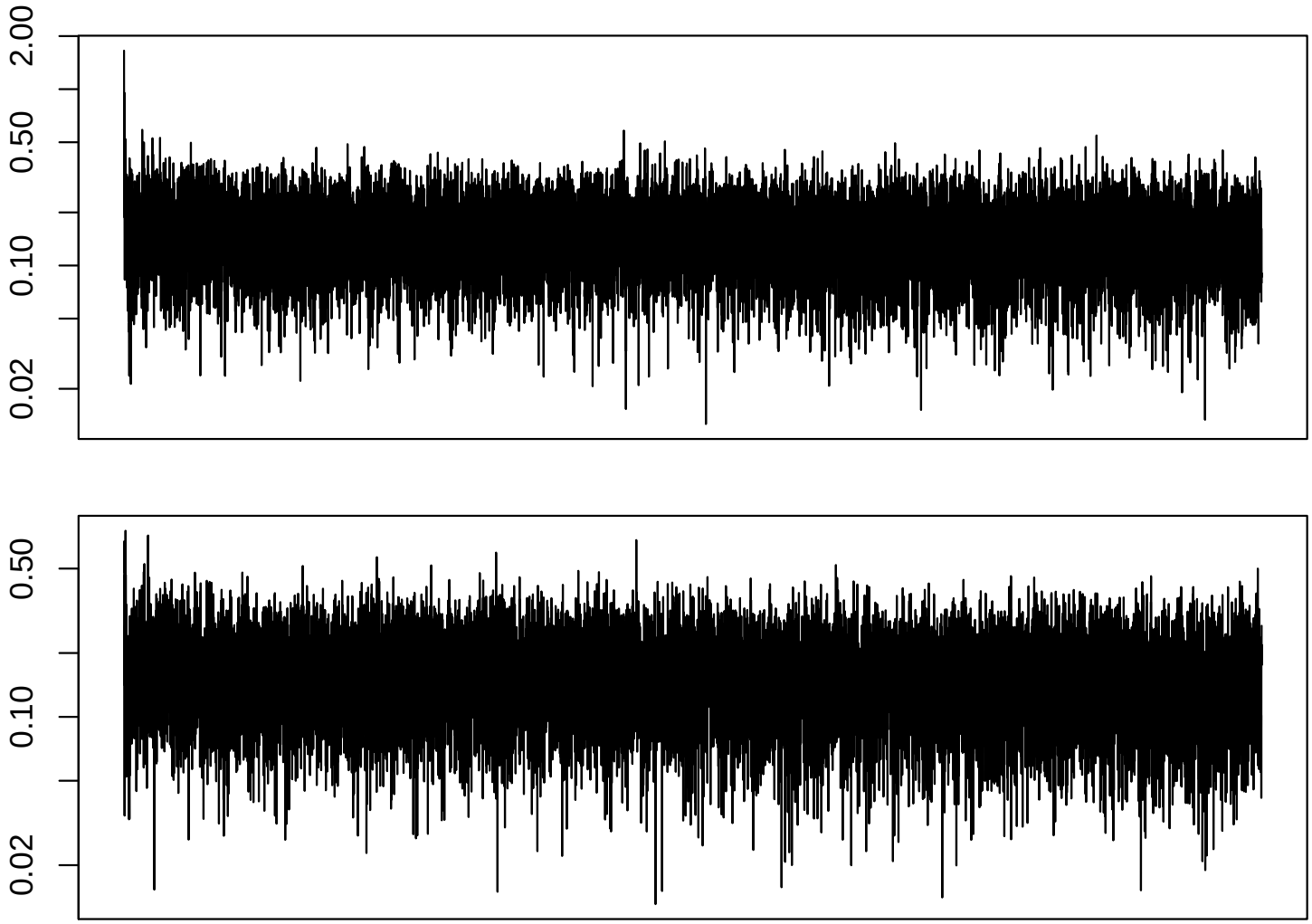
Effect much smaller for easterly

Spectra of innovations



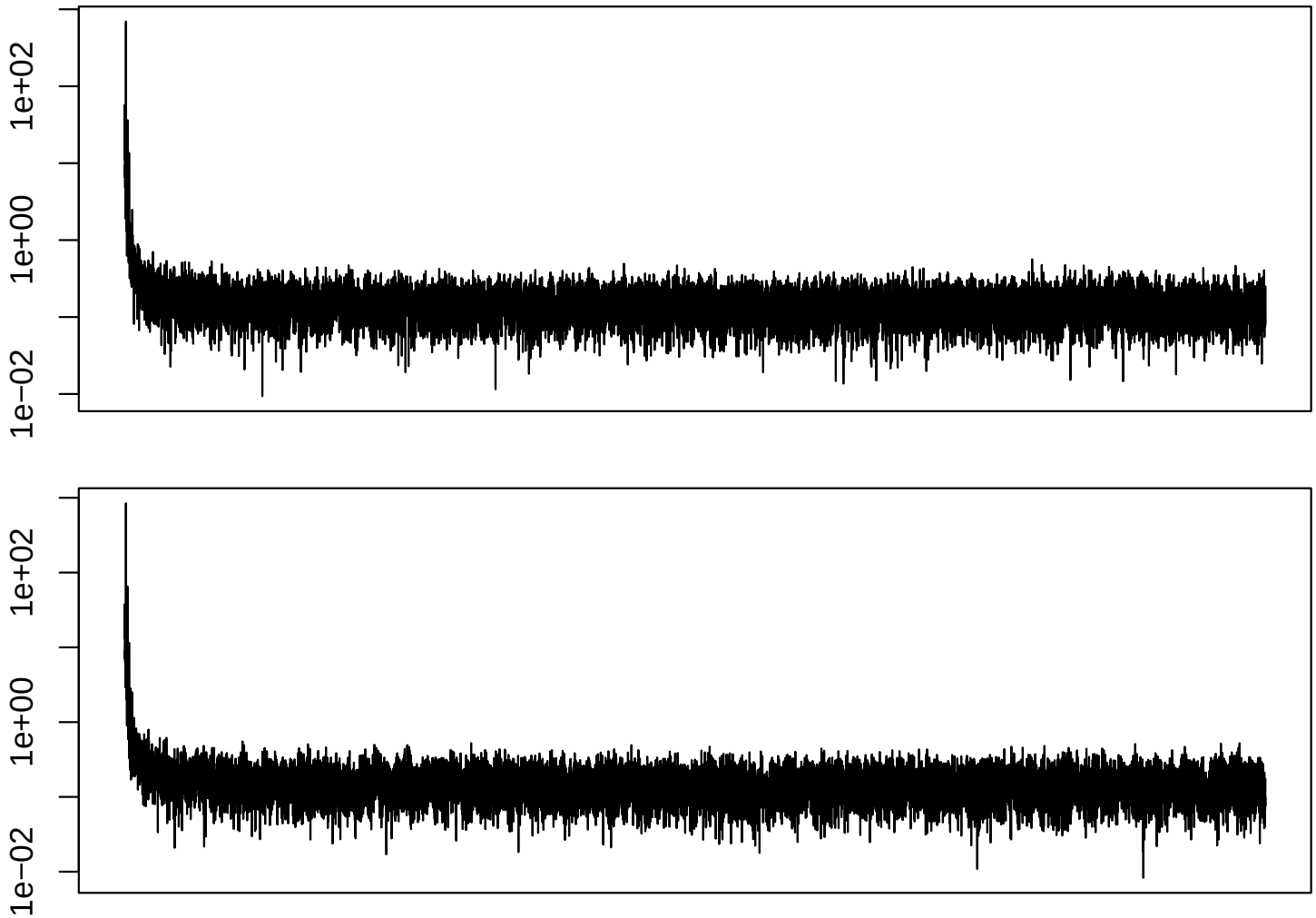
Spectra of innovations (upper = northerly, lower = easterly) averaged over the 5 years

Spectra of absolute innovations



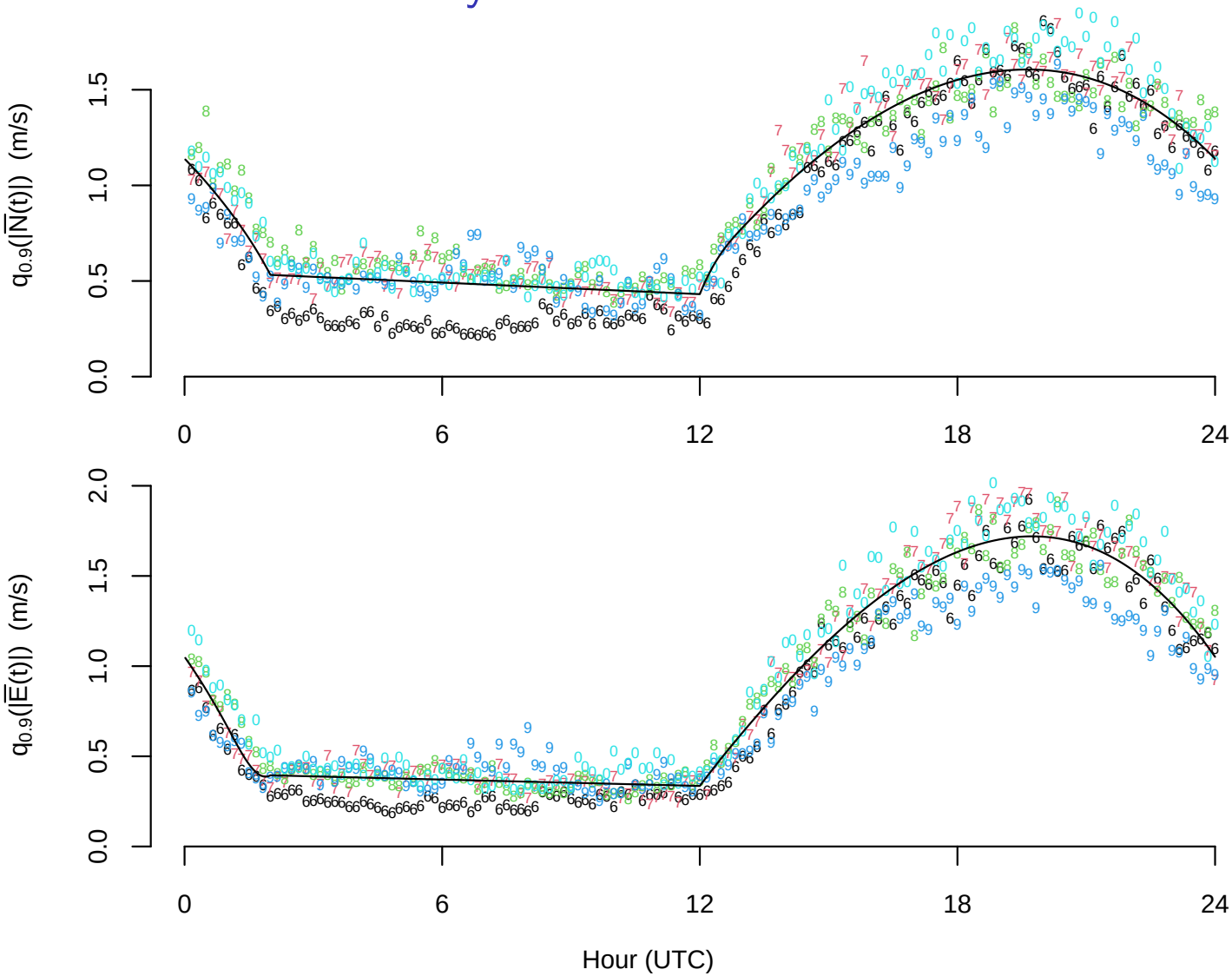
Spectra of absolute innovations (upper = northerly, lower = easterly) averaged over the 5 years

Spectra of absolute residuals correcting for diurnal but not stochastic variation



Spectra of absolute diurnally normalized residuals (upper = northerly, lower = easterly) averaged over the 5 years

Unconditional diurnal volatility



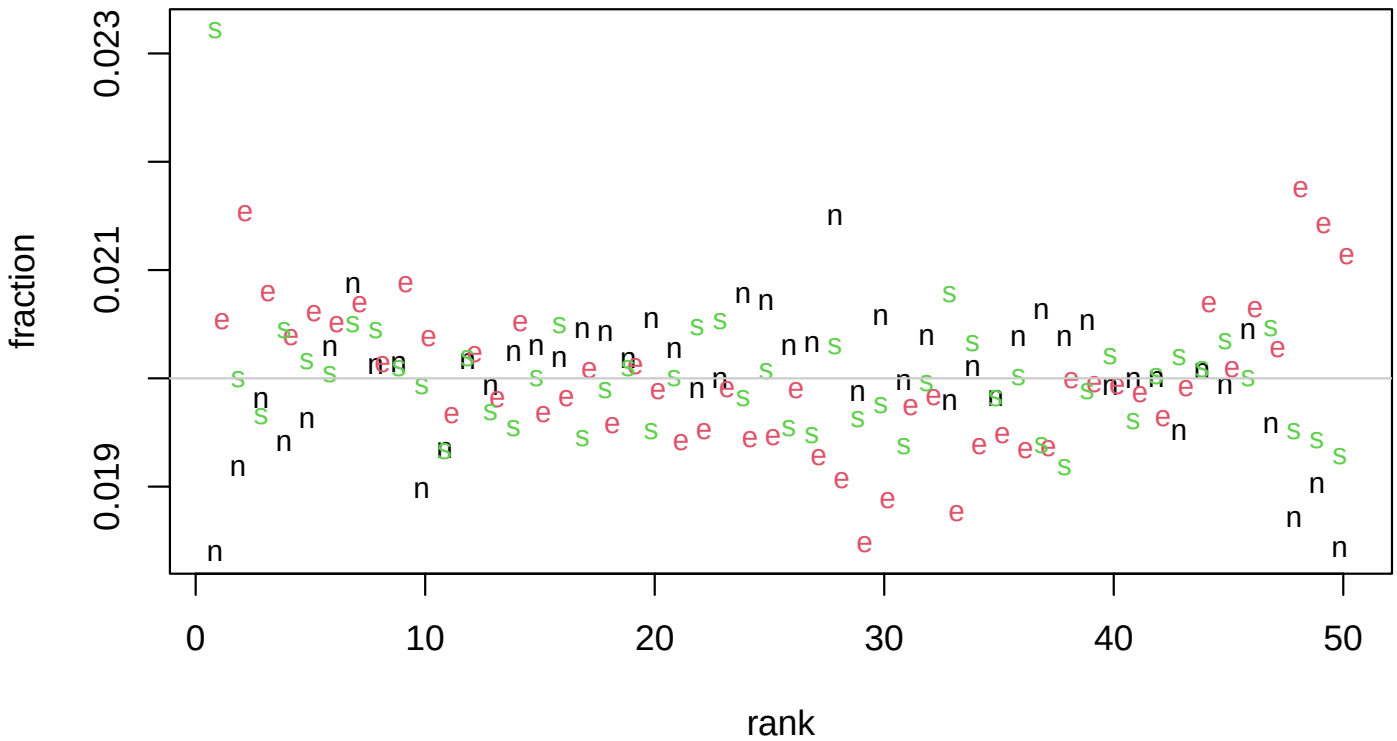
Empirical 0.9 quantiles of $|\bar{N}(t)|$ (upper) and $|\bar{E}(t)|$ (lower) for 10-minute periods by year

- Black curves are fitted $DV_N(m)$ and $DV_E(t)$

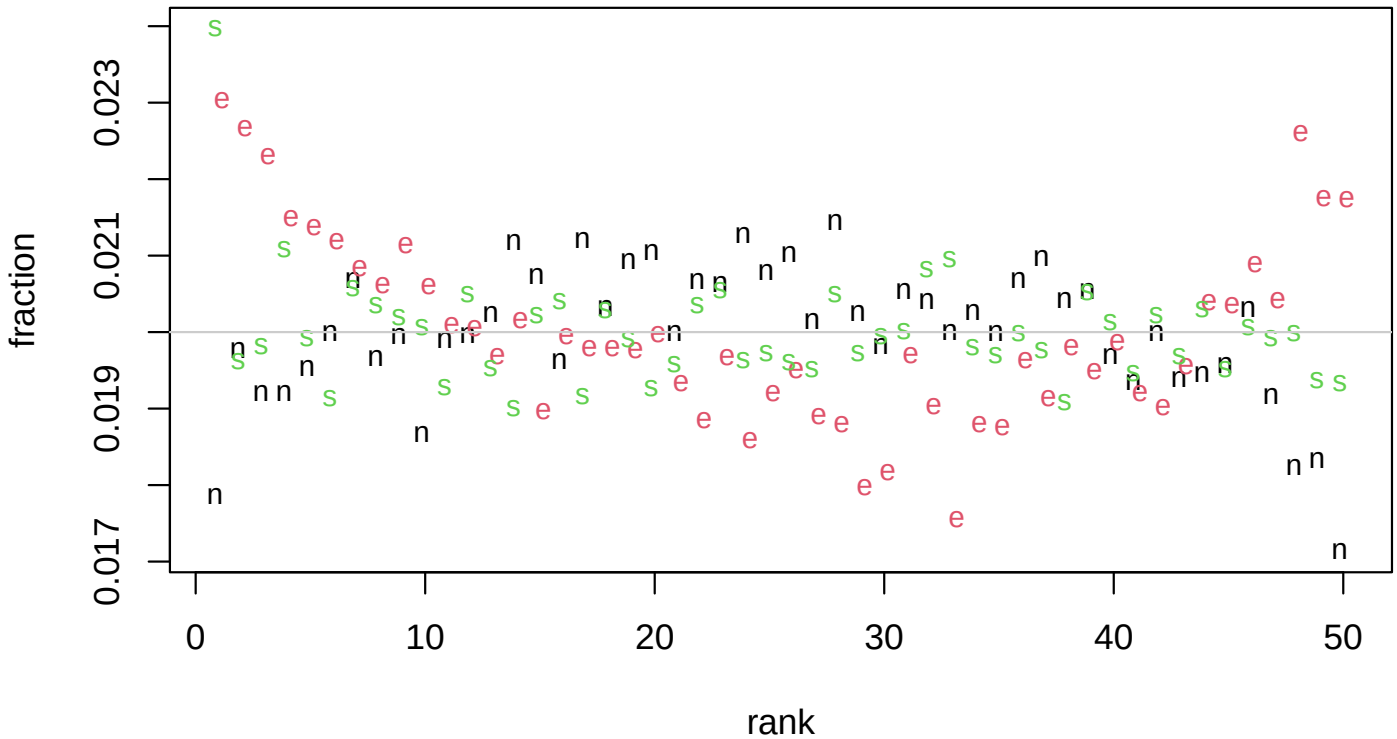
Calibration of one-minute ahead predictions

For each minute, simulated 49 one-minute ahead predictions

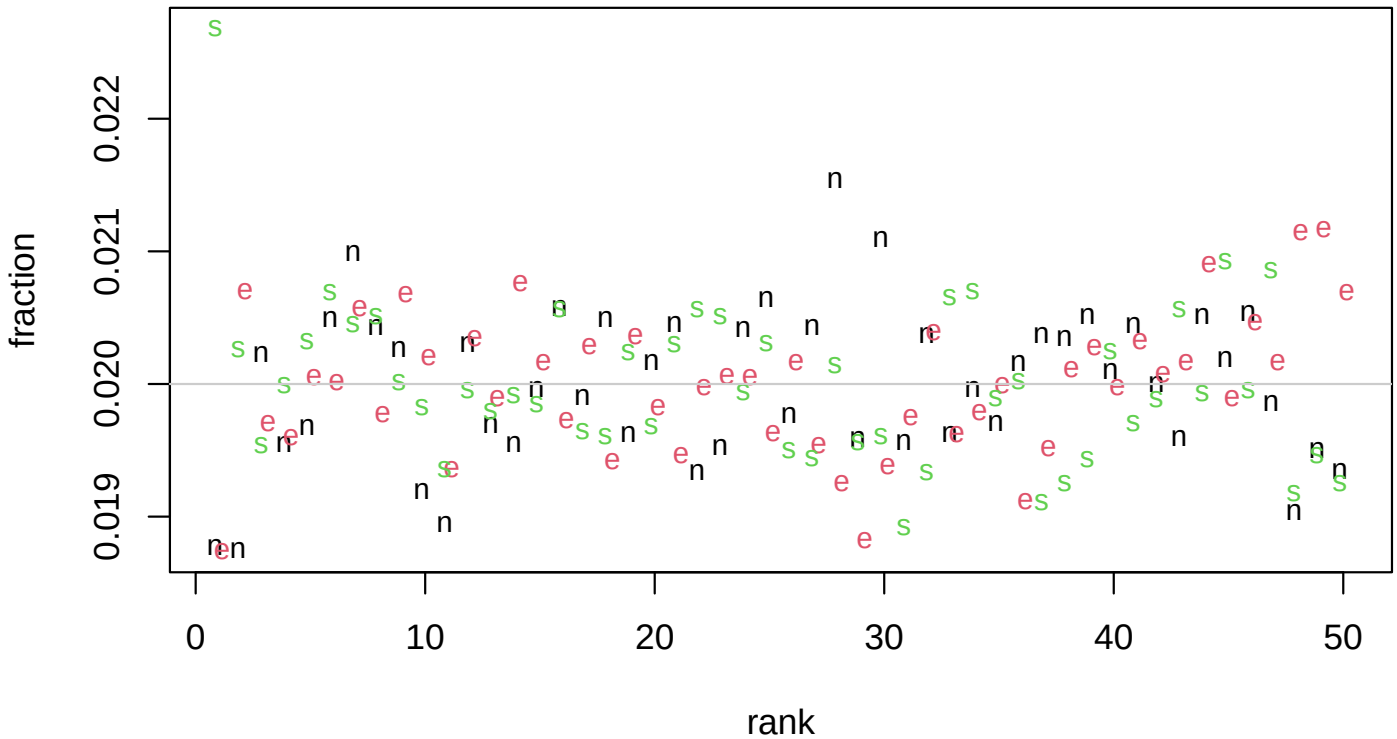
- Among simulations, actual wind vectors should have equal probability for ranks 1–50



Ranks of observed winds among 49 one-minute ahead predictions
n = northerly, e = easterly, s = speed



Same as previous plot for nighttime hours



Same as previous plot for daytime hours

Compare observed and simulated winds

Simulations are fairly fast. I have simulated 25 years, but could easily do more.

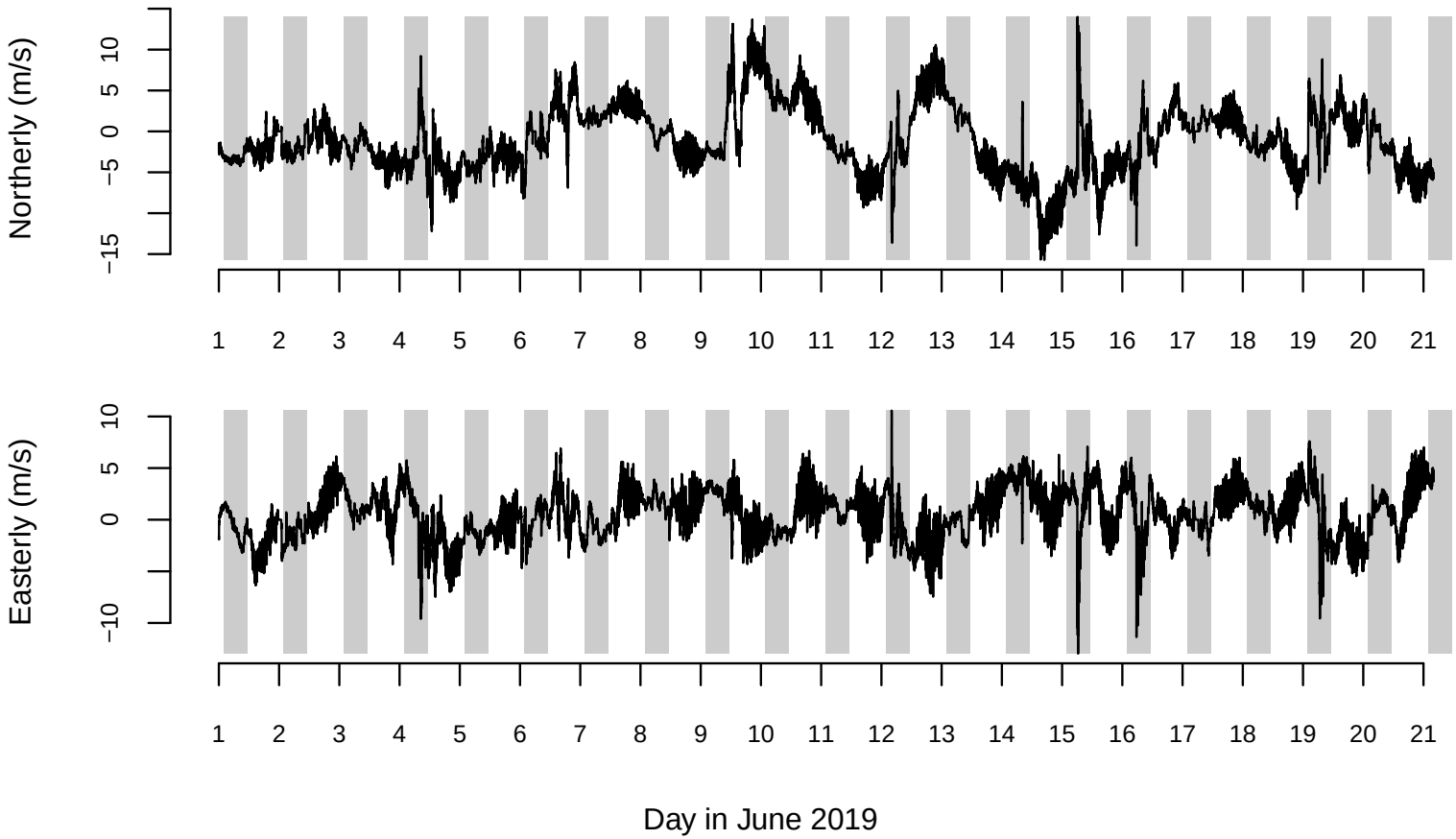
- Use 4 hours of real data to get started, but could use random starting points and burn-in period
- Following plots reshow initial ones for data and compare to simulations

Overall, agreement is pretty good, but

- I doubt my model would produce a simulation like 2020
  - Need (non-Markov) regime switching model?
- Produces winds higher than observed somewhat too often?
- First difference at night look off
  - Problem with rotation?
  - If don't rotate, simulations produce unrealistic wind speeds
  - Maybe only sometimes rotate, but haven't yet found a good way to do this

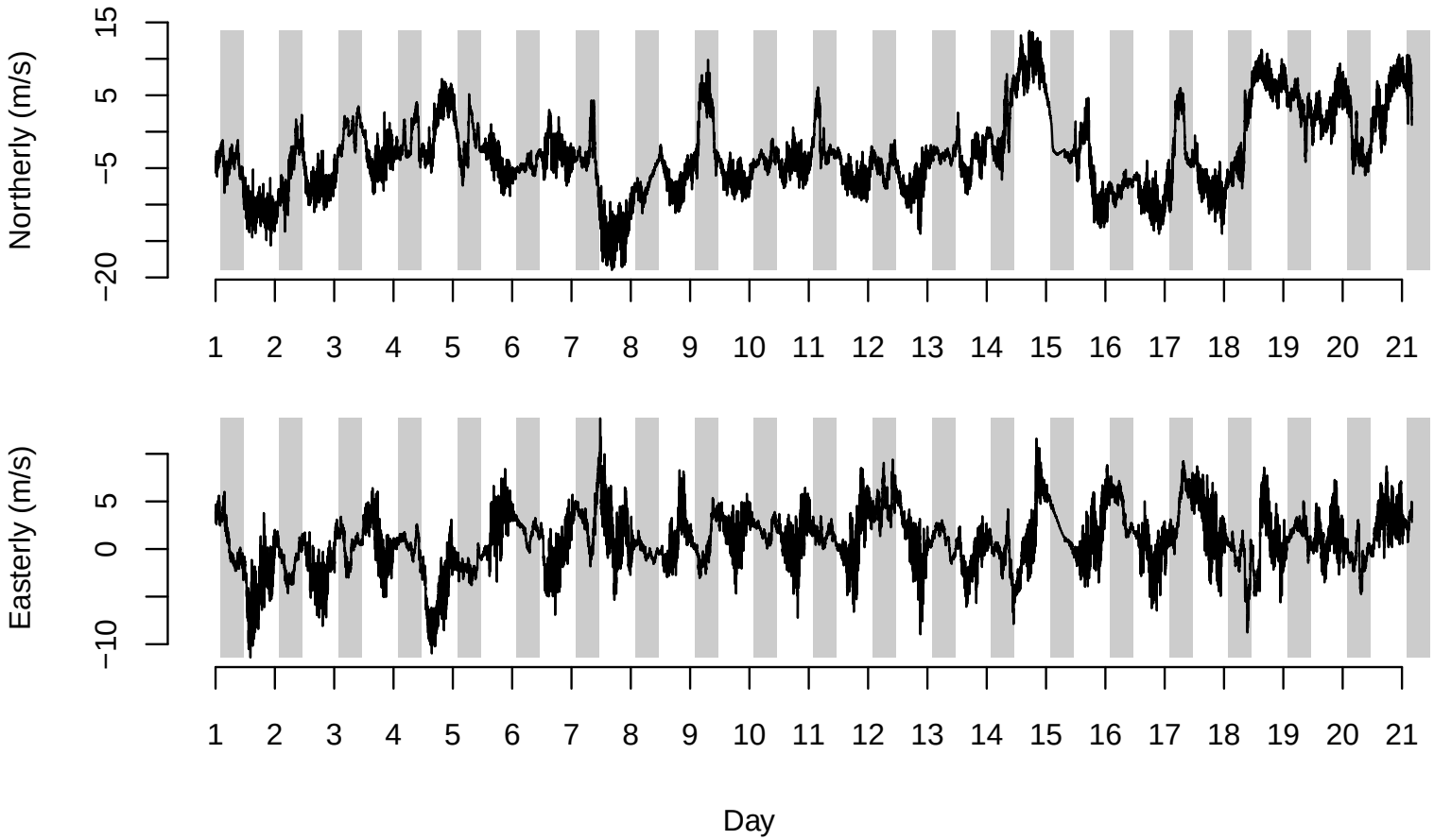
Other models do better in some regards, but worse in others

Observed



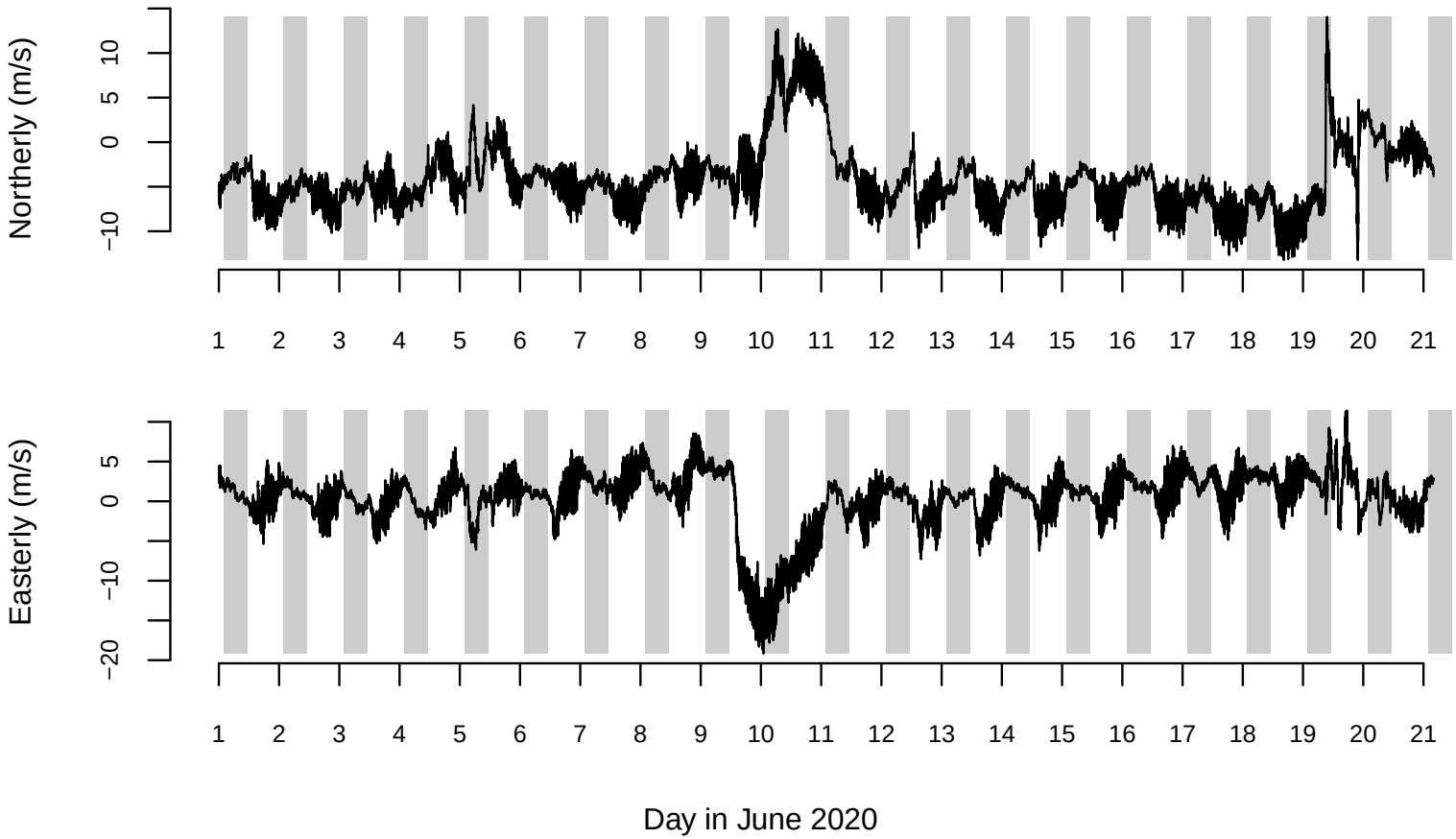
Wind vectors, 2019. Gray = nighttime.

Simulated



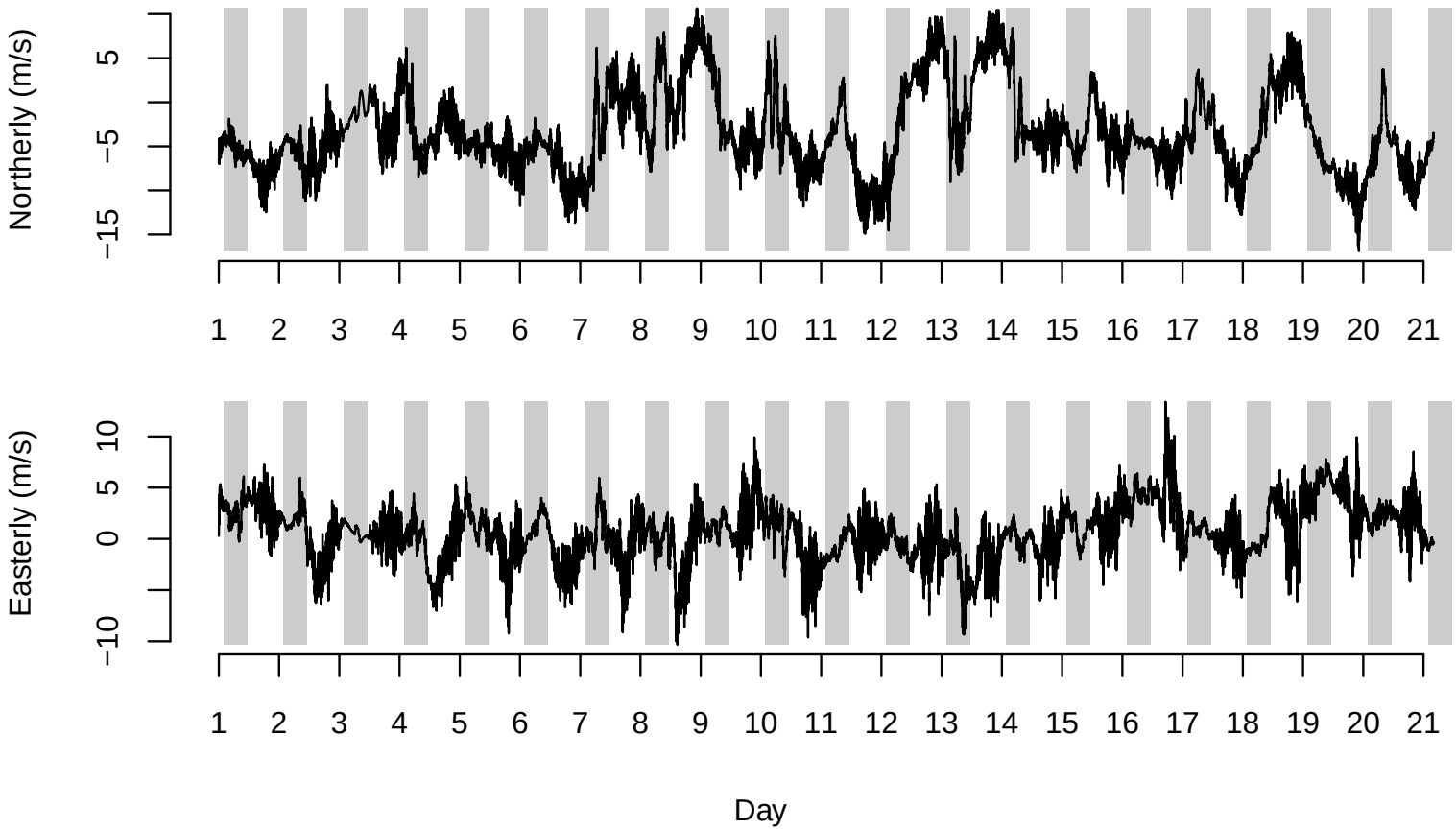
Wind vectors, June 21, 2019 starting values. Gray = nighttime.

Observed



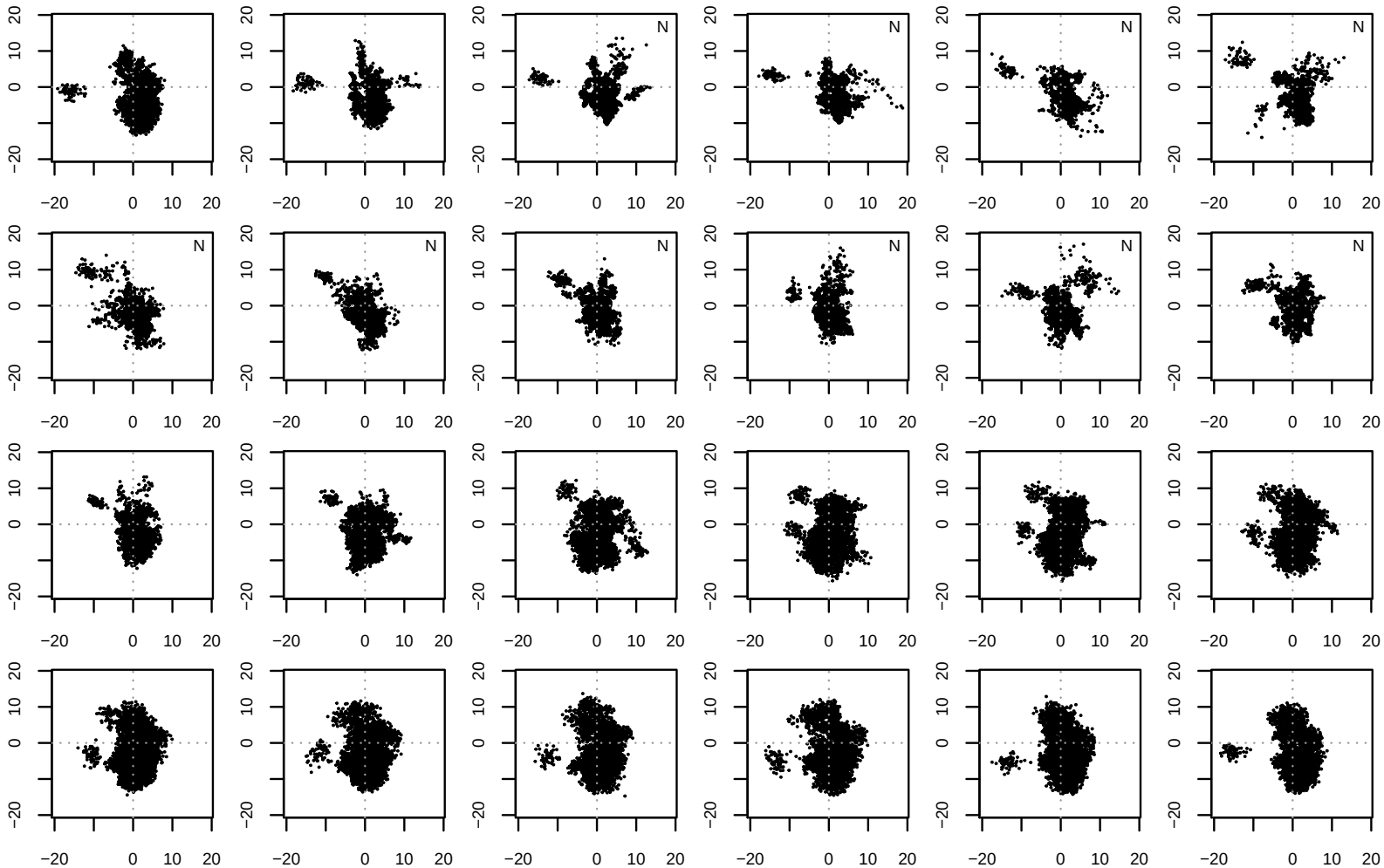
Wind vectors, 2020. Gray = nighttime.

Simulated



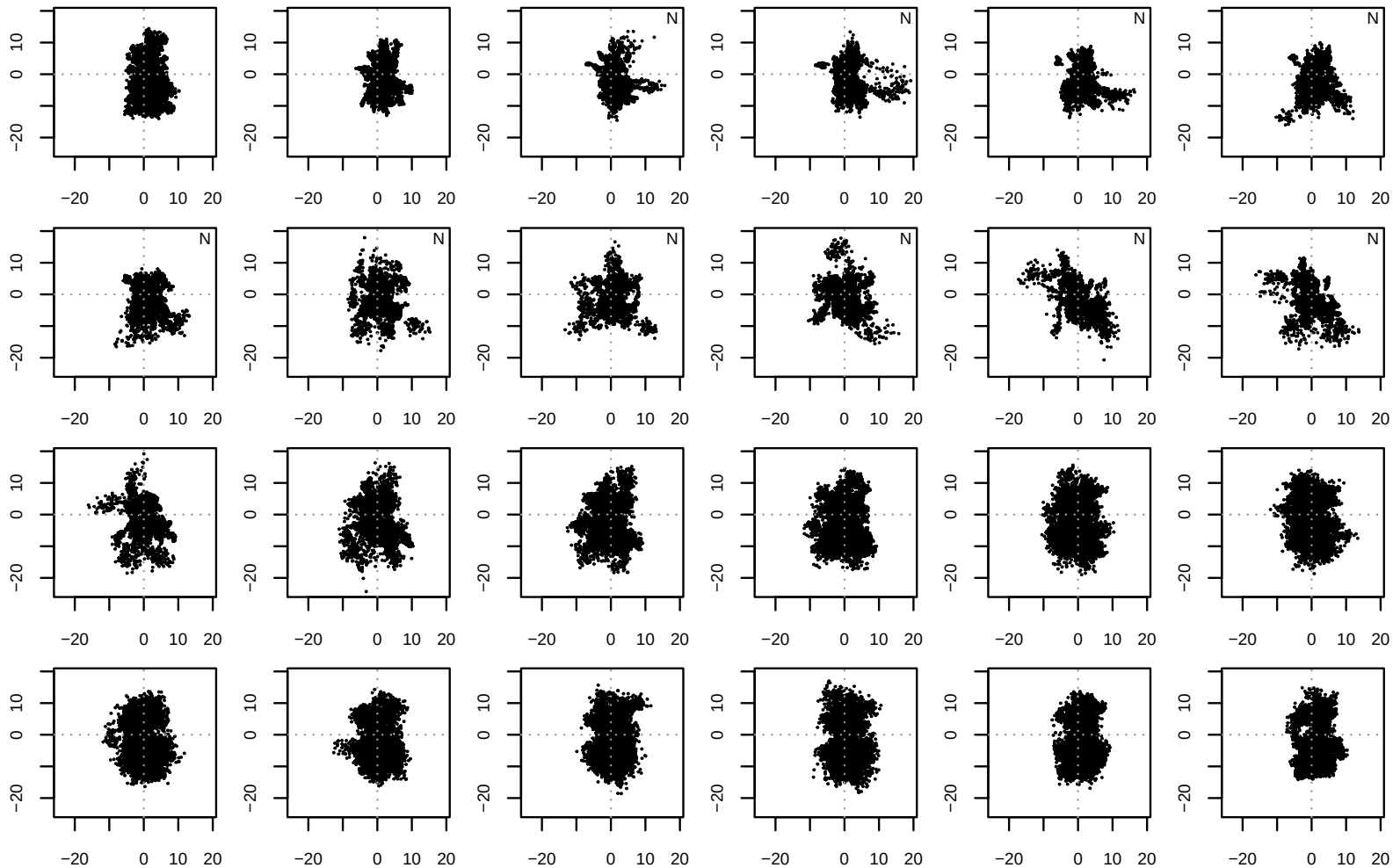
Wind vectors, June 21, 2020 starting values. Gray = nighttime.

Observed



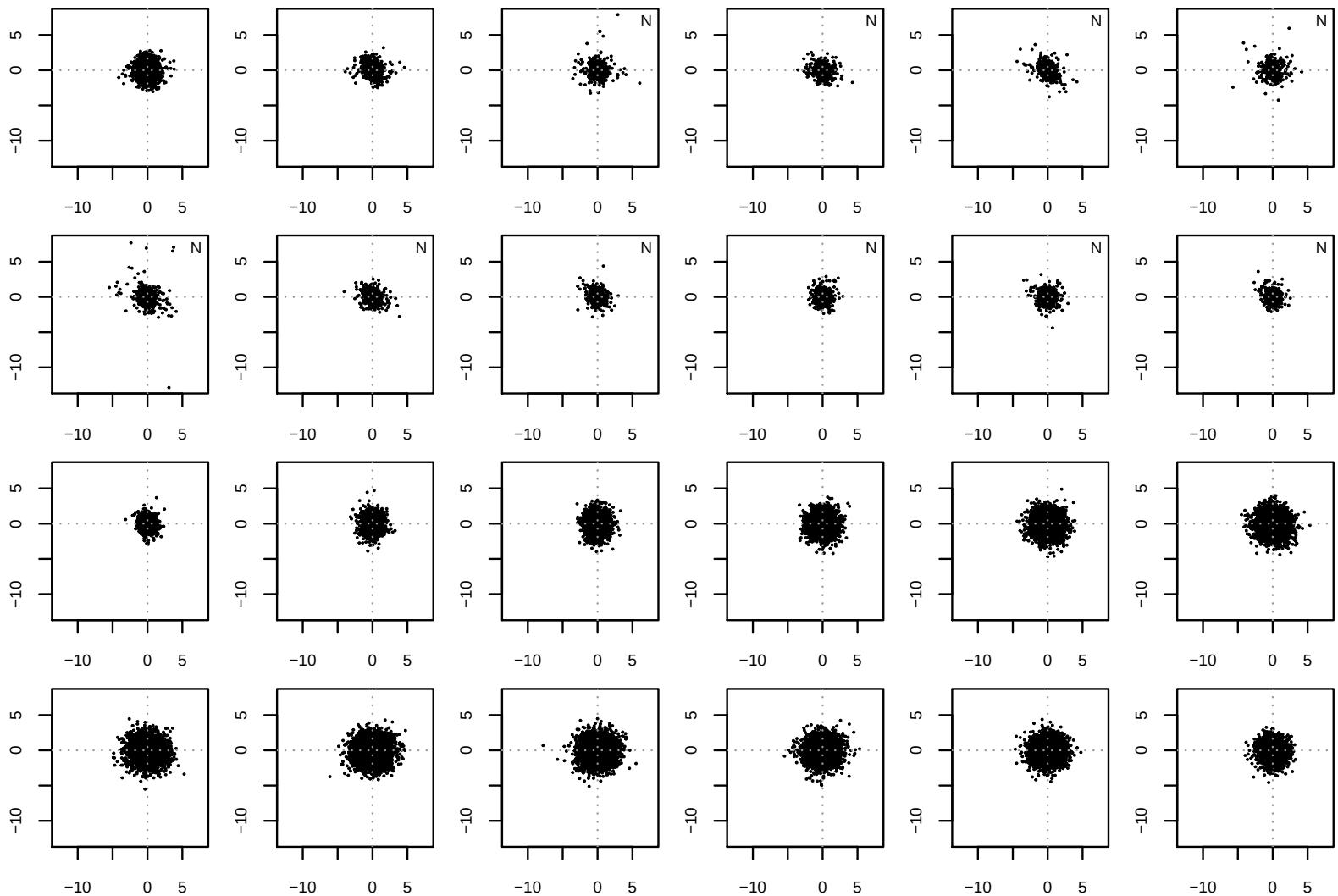
Wind vector (m/s) by hour (UTC).

Simulated



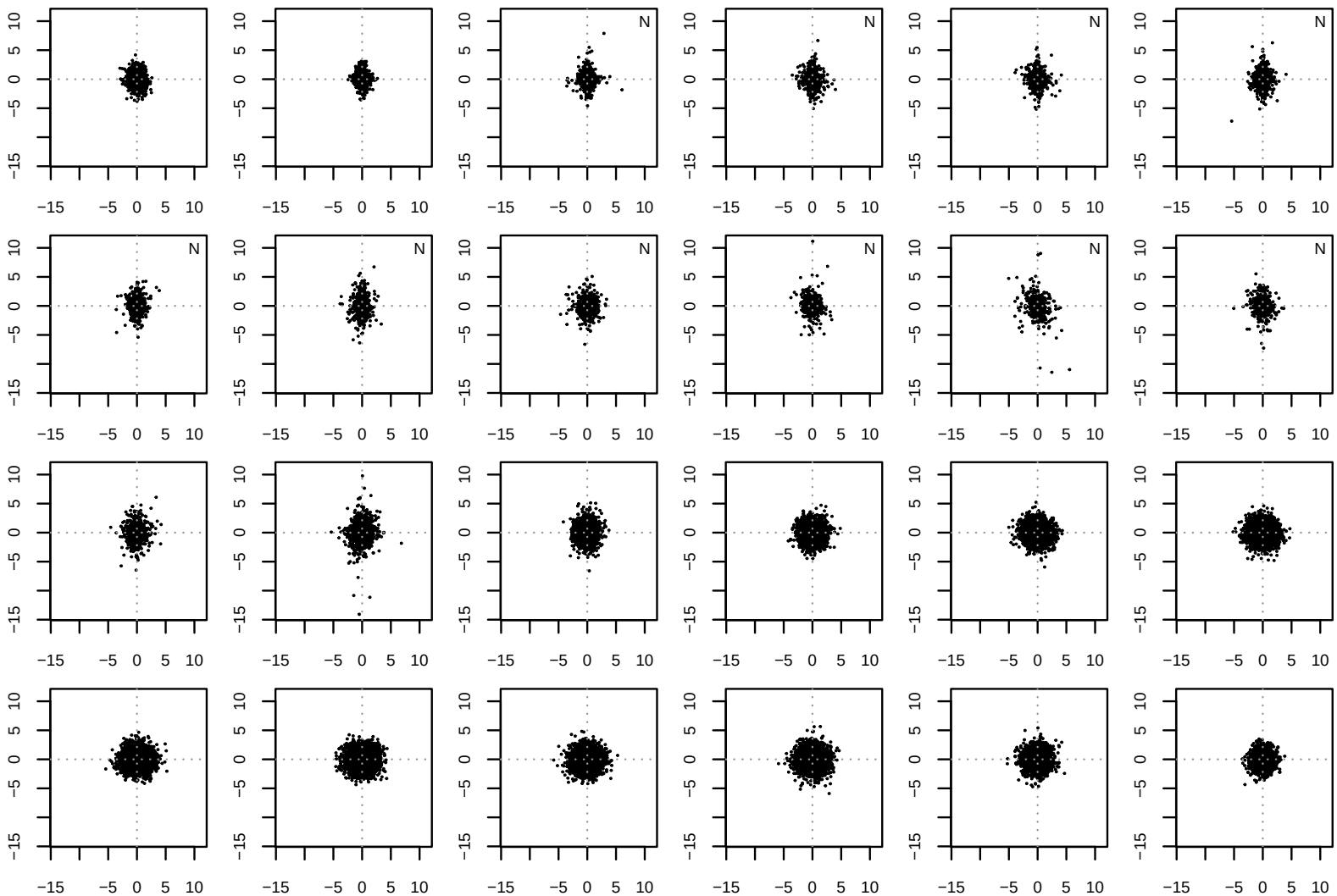
Wind vector (m/s) by hour (UTC).

Observed



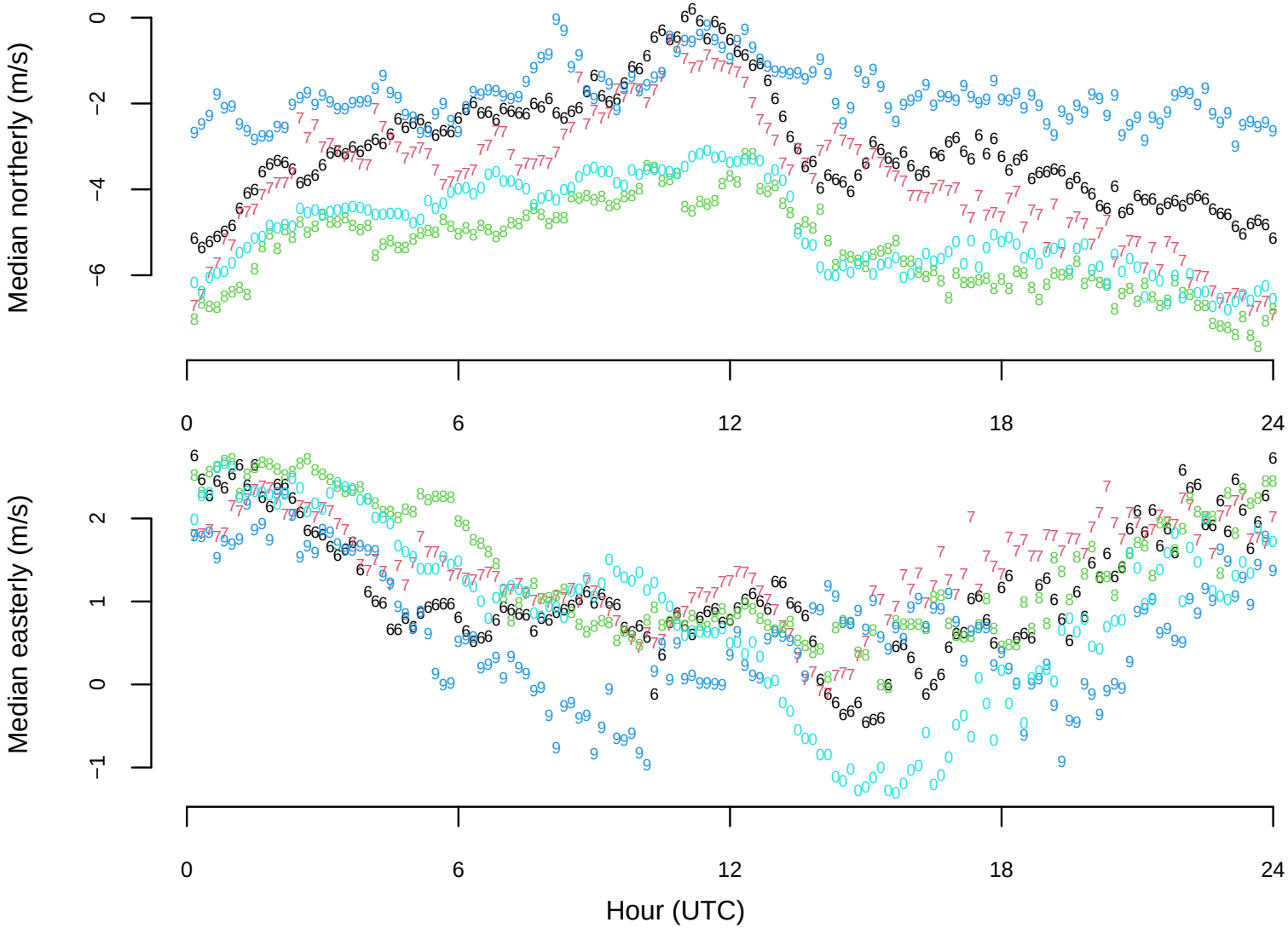
First differences in wind vector (m/s) by hour.

Simulated



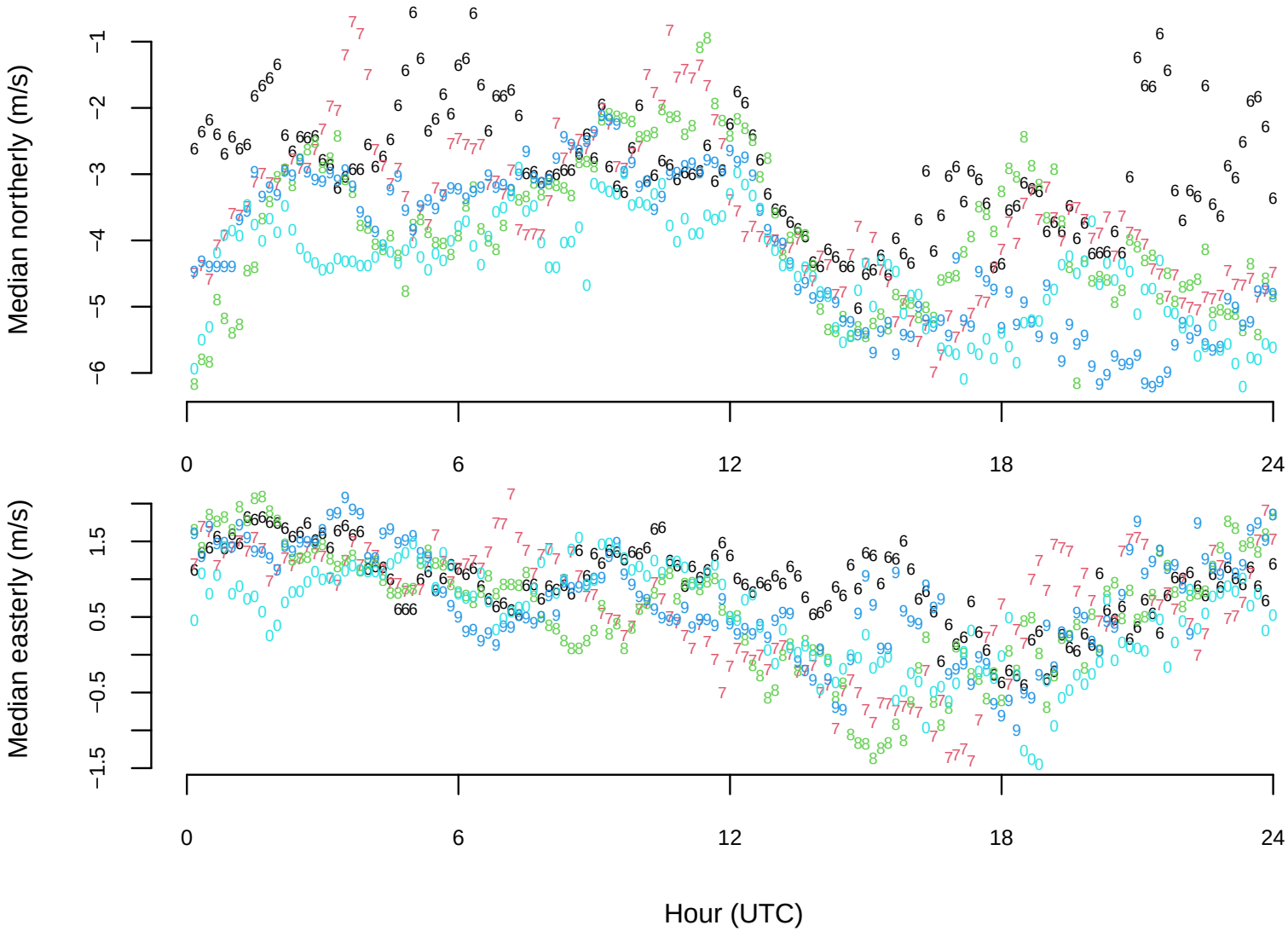
First differences in wind vector (m/s) by hour (axes not same as previous plot).

Observed



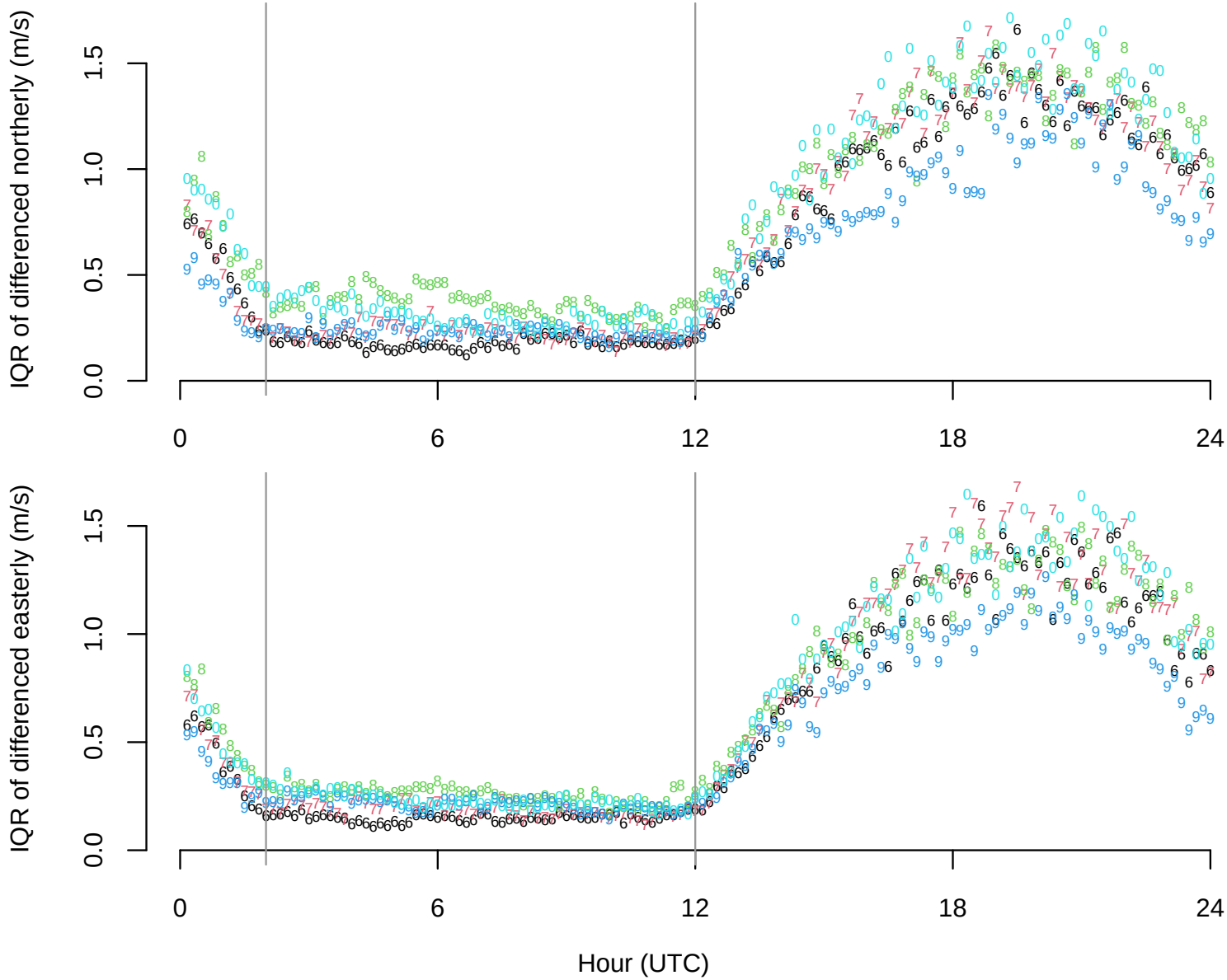
Top: Medians of one-minute northerly wind for every 10-minute period. Plotting symbol = last digit of the year. Bottom: Same for easterly component.

Simulated



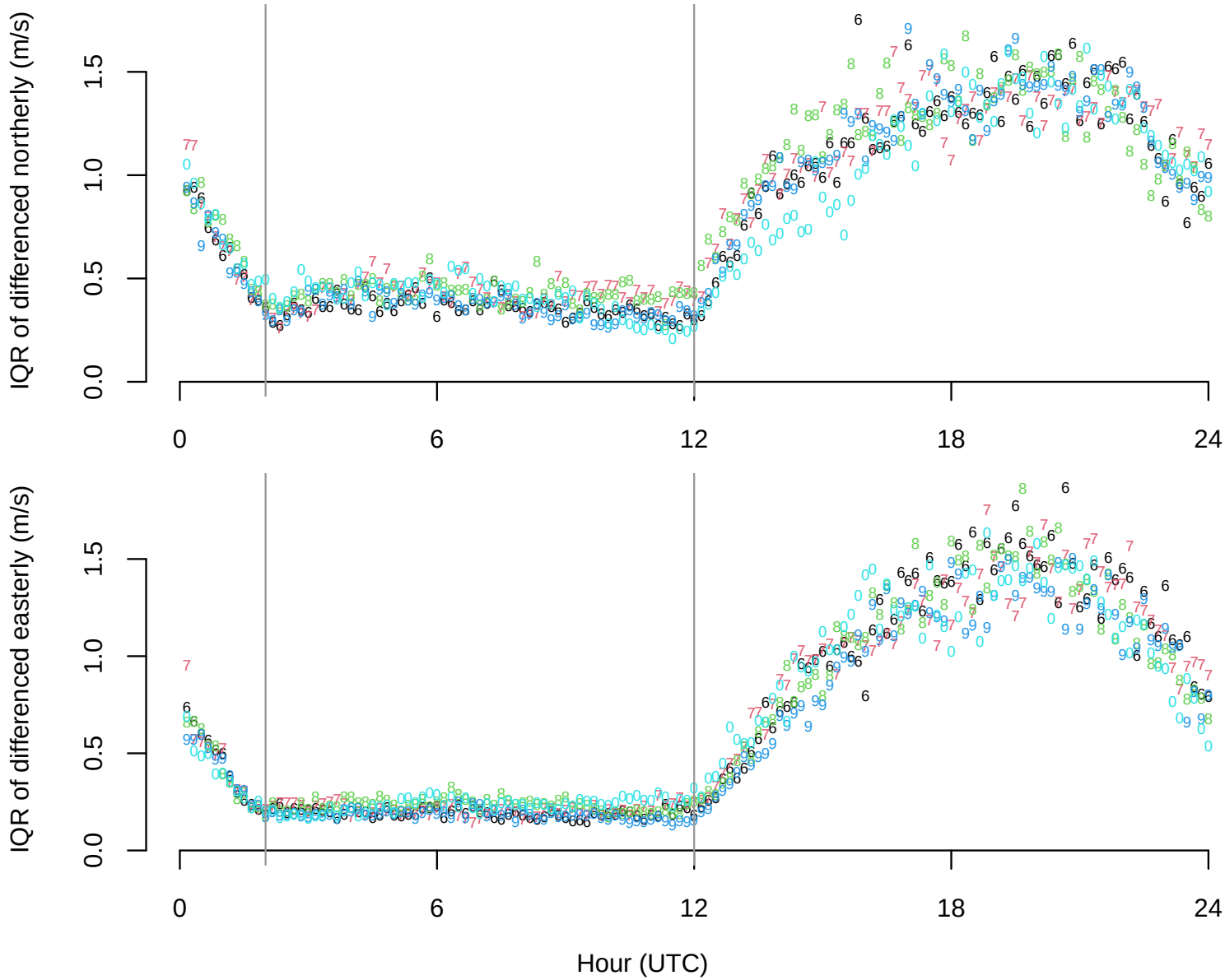
Top: Medians of one-minute northerly wind for every 10-minute period. Plotting symbol = last digit of the year. Bottom: Same for easterly component.

Observed



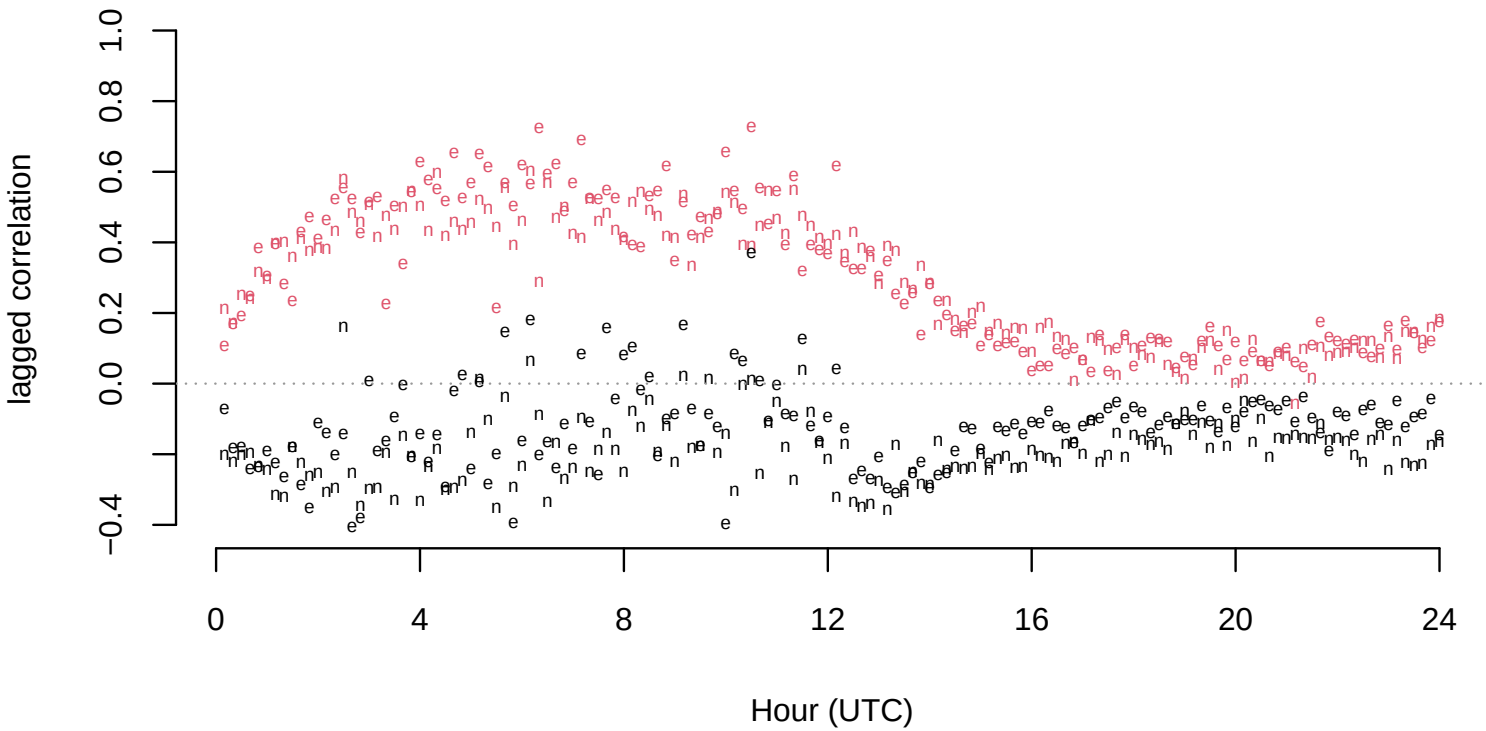
Interquartile range of first differences. Gray lines delineate nighttime.

Simulated



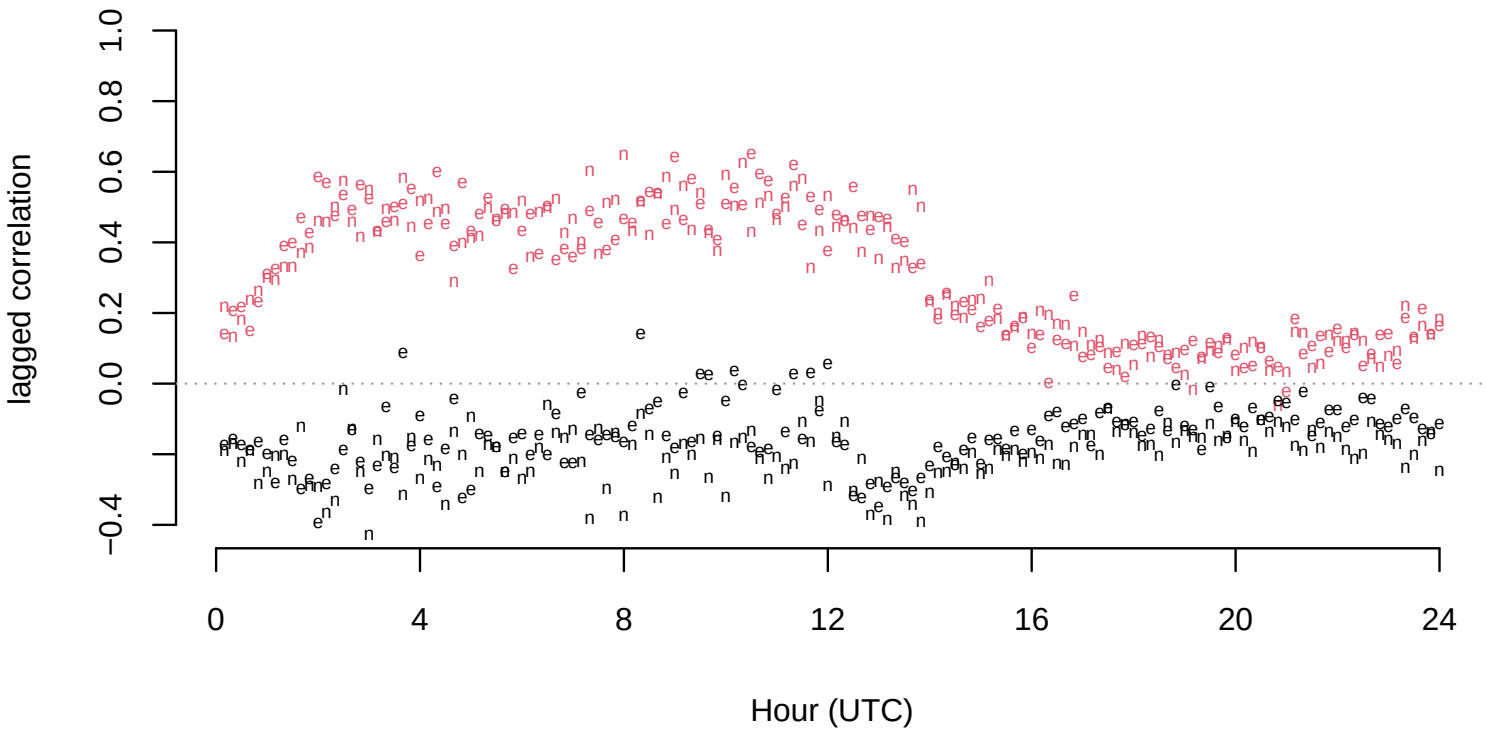
Interquartile range of first differences. Gray lines delineate nighttime.

Observed



Correlations in first differences of wind vector (black) and absolute values of first differences (red). n = northerly component and e = easterly.

Simulated



Correlations in first differences of wind vector (black) and absolute values of first differences (red). n = northerly component and e = easterly.

Uncertainty quantification

No formal hypothesis tests

No confidence intervals or even standard errors

Why not? Suppose condition on chosen model

- Still not clear how you would obtain appropriate standard errors since innovations not iid
- If had used more years, maybe resampling years would be reasonable for inferences on conditional median and spread?
- UQ for sampling approach to innovation distribution?

But surely the extensive fishing for a model affects uncertainties

Best bet would be to look at other years of data, but I am not yet ready to cross that barrier

Comparisons to machine learning approach

This work has required *lots* of my time

- I have fit many separate models for conditional medians and spreads

Includes several changes of basic form

- How to capture unconditional diurnal cycle in median
- How to ensure conditional spread is positive
- Inclusion of nonlinear effects

I have fussed the most with modeling innovations

- Started with bivariate t with diurnally varying degrees of freedom and correlation
- Moved to simple sampling scheme of just resampling empirical innovations from same time of day

Could a machine learning approach (e.g., recurrent neural net) do

- almost/equally as well
- even better

with much less human effort?

Possibly, but

- I am skeptical it would produce a good simulator if trained solely on one-minute ahead predictions

Could consider forecasting at multiple time lags (I did some of this)

- Recursive nonlinear prediction based on one-step ahead predictions requires model for bivariate conditional density

What existing software would be adequate?

- Many programs focus on point prediction
- Would any of them get diurnal patterns accurately without help?

If only goal were to produce a stochastic weather generator for horizontal wind vector in Lamont, OK, then I have wasted my time

Hope is to gain understanding about

- Meteorologically interesting patterns in high-frequency horizontal wind vectors that apply more generally, e.g.,
  - diurnal cycles, stochastic volatility, nonlinearities
- Structures that one may want to include in statistical models for other high-frequency environmental time series

Understanding needs interpretable models

I suspect that a special purpose machine learning approach could work quite well

- Especially if some of the insights I have gained were incorporated
- What level of coding effort would this require?
- Would resulting model be interpretable?

There is a lot more data

- What if gave algorithm the full 29 years of wind vectors (about 15 million time points)?
- What if gave algorithm other surface meteorology?
  - Should help for prediction, but would it help for simulation?

I am unqualified to take this on

- Anyone else?

Things left out

Need for dependence on longer time scales?

- Recurring patterns across several days (2020)
- Differing patterns across years
  - Allowing some parameters to vary across years didn't obviously help
- Regime-switching model?

Seasonality

- Interaction with diurnality must be substantial
- Basis functions used should depend on time of sunrise/sunset so need to vary with season

Extremes

- Wrong data for this purpose
- No tornadoes in these data. What are chances one hit this station over last 29 years?

Not to mention spatial variation in horizontal/vertical

Take-home messages

Whenever I look carefully at real environmental data, I essentially always find existing models are inadequate

- Surely the same holds for data from most (all?) areas
- Therefore, the need for good diagnostics is paramount
- For processes with complex dependencies, simulating from fitted model provides a particularly challenging test

Labor intensive model development will remain a valuable part of the applied statistician's toolkit

But maybe newer methods can improve or at least speed up the process?