# Inferential Challenges for Large Spatio-Temporal Data Structures

# Day 2
## *Large scale models*

# Today's schedule (morning)

- 9:00-9:10 *Wit, Insight, and Matters of Great Importance* [T. Duchesne]
- 9:10-9:55 **Regional climate model assessment via spatio-temporal modeling** [P. Craigmile]
- 9:55-10:05 Discussant [D. Becker]
- 10:05-10:15 Floor discussion

- 10:15-10:45 Coffee Break



- 10:45-11:30 **A projection-based approach for spatial generalized linear mixed models** [M. Haran]
- 11:30-11:40 Discussant [J.-F. Coeurjolly]
- 11:40-11:50 Floor discussion

- 11:50-12:00 *Interesting and important stuff* [T. Duchesne]

- 12:00-**13.15** Lunch

# Today's schedule (afternoon)

- **13:15**-13:35 **GROUP PHOTO**

- 13:35-14:20 **Distributed Spatial Kriging: Scalable Bayesian Framework for Massive Spatially Indexed Datasets** [R. Guhaniyogi]
- 14:20-14:30 Discussant [B. Taylor]
- 14:30-14:40 Floor discussion

- 14:40-15:15 Coffee break

- 15:15-16:00 **Challenges in modelling geolocated health data** [T. Smith]
- 16:00-16:10 Discussants [L. Waller, P. Nguyen]
- 16:10-16:30 Floor discussion

- 16:30-16:45 *Wit, Insight, and Matters of Great Importance* [T. Duchesne]

- 17:30-19:30 Dinner

# What does "large scale" mean?

My definition of a large scale situation:

**When what you usually do and works well stops working well because something has become too large.**

# Challenges with large scale models

The obvious one : **computational issues** (i.e., data storage, memory, number of floating point operations, etc.)

- inversion of large, <span style="color:red">**non-sparse**</span>, matrices

- numerical integration in high-dimension

- interpolating/predicting the value of this process at (potentially a large number of) new locations!

# Challenges with large scale models

The obvious one : **computational issues**

Examples that we will see today:

- Large correlation/covariance matrices that must be inverted
- Models with a large number of random effects that must be integrated out
- High dimensional posterior distributions computation (or simulation from)
- Comparing output of model(s) to observations over large spatial domain
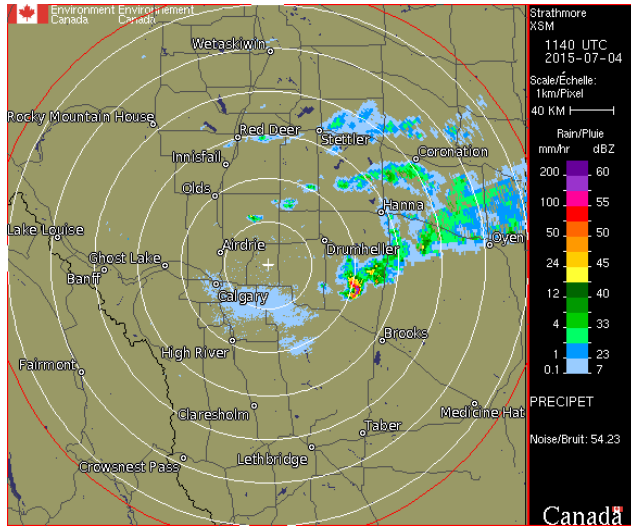
# Challenges with large scale models

The obvious one : **computational issues**

Example of another type of computational challenge:

- Integrating a large number of Doppler radar images over time to improve estimation of cumulative precipitation [would be very useful for forest fire management agencies]

# Challenges with large scale models
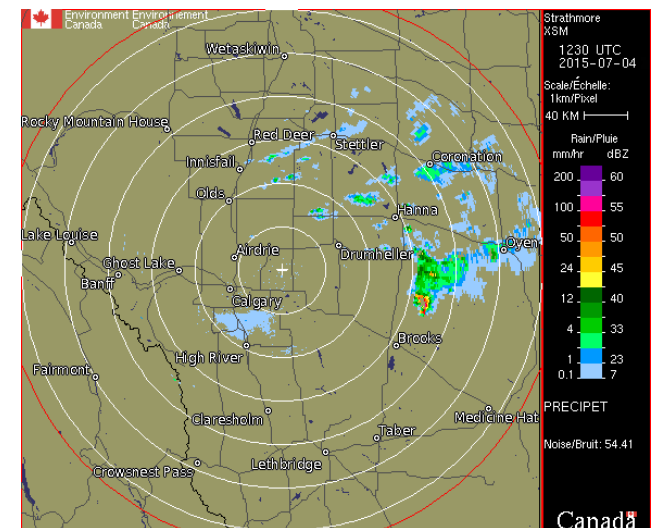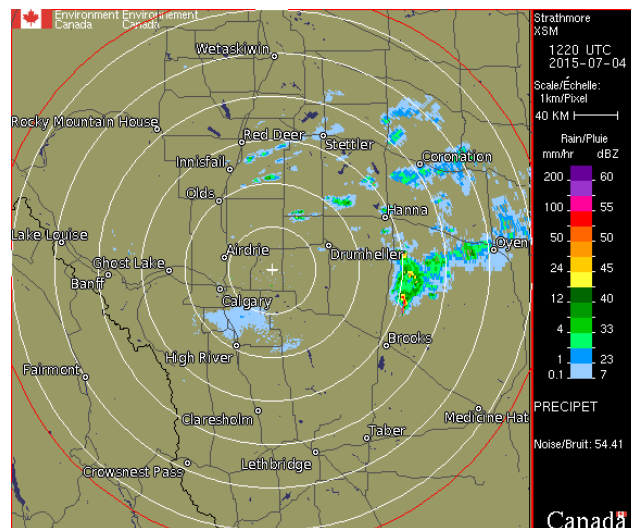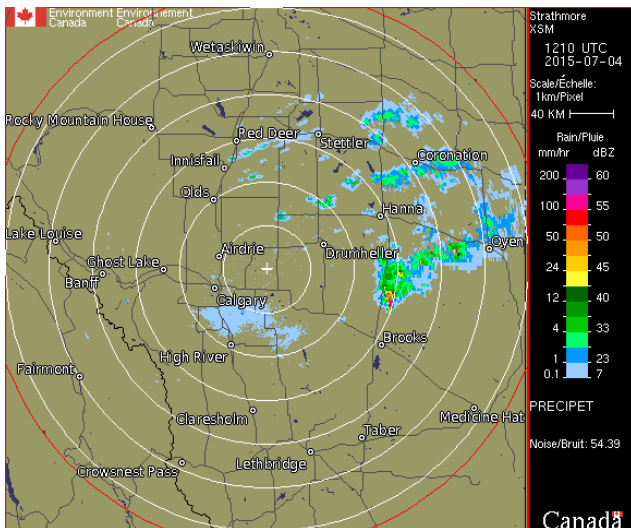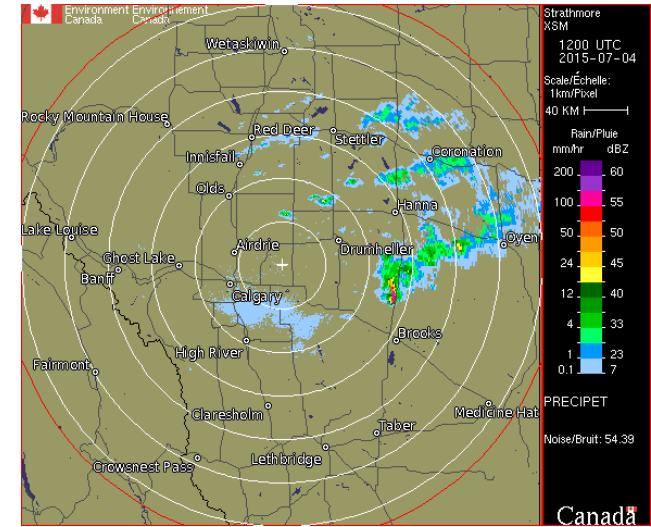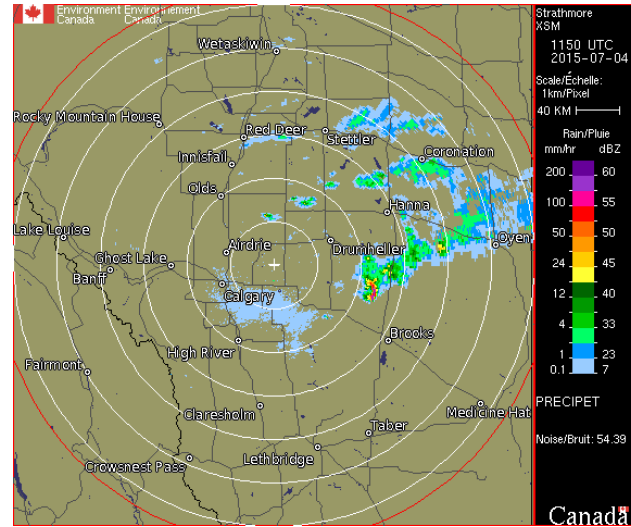
# Challenges with large scale models

NEXRAD on AWS

*The [Next Generation Weather Radar](#) (NEXRAD) is a network of 160 high-resolution Doppler radar sites that detects precipitation and atmospheric movement and disseminates data in approximately 5 minute intervals from each site. NEXRAD enables severe storm prediction and is used by researchers and commercial enterprises to study and address the impact of weather across multiple sectors.*

***The real-time feed and full historical archive of original resolution (Level II) NEXRAD data, from June 1991 to present, is now freely available on Amazon S3 for anyone to use****. This is the first time the full NEXRAD Level II archive has been accessible to the public on demand. Now anyone can use the data on-demand in the cloud without worrying about storage costs and download time.*

# Challenges with large scale models

The less obvious one : **modeling issues**

- Difficult to obtain a valid joint distribution over all *n* observations when global model is broken into several smaller pieces
- Priors must be specified for a large number of parameters … which translates into several hyperparameters to specify/model
- Stationarity, isotropy and other simplifying assumptions tend to hold locally, but not over a large scale

# Challenges with large scale models

The less obvious one : **modeling issues**

Another example in property insurance:

- Can a company come up with a single geographic rating model for a given jurisdiction (e.g., province), which consists of the union of a large number of highly heterogeneous areas (e.g., large cities, small cities, rural areas, cottage country, etc.)?

# Challenges with large scale models

Another less obvious one : **inferential issues**

- Spatial confounding

- Predictors and response measured at different scales or locations, or different predictors measured in different areas

- Non-uniform non-detection

# Challenges with large scale models

Solutions proposed today

- Compare distributions of observations and large-scale model output
    > Use key features of these distributions

    > Approximations
- Reduce the dimensionality of random effects
    > PCA using random projections
- Divide-and-conquer
    > Don't break-up space into pieces but use K representative sub-samples
- Fit full spatio-temporal latent GP
    > Use extended grids, fast Fourier transform, additivity assumptions
- Address spatial confounding
    > Modeling
    > Restrict random effects to be orthogonal to fixed effects