ガウス過程のシミュレーションへの応用
～季節変動を考慮した動的台風移動モデルの構築～

講師: 斎藤正也

資料提供: 中野慎也

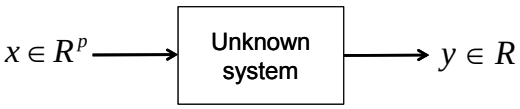
(情報・システム研究機構 統計数理研究所)

目次

- シミュレーションへのGPの利用
- GPの復習
- 事例: 季節変動を考慮した動的台風移動モデルの構築
  (統計数理研究所　中野慎也 助教)

GPによる回帰

- 回帰問題: $y = f(x) + \varepsilon, \quad \varepsilon \sim \mathbf{N}(0, \sigma^2)$



- GP回帰:

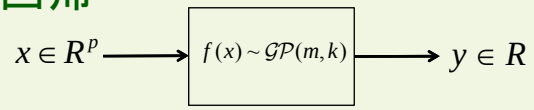
$$\mathbf{f} = [f(x_1), f(x_2), ..., f(x_n)]_{n=1,...,\infty}, \quad \mathbf{f} \mid \mathbf{x} \sim \mathbf{N}(\mathbf{m}, \mathbf{K})$$

$$f(x) \sim \mathcal{GP}(m(x), k(x, x'))$$

$$m(x) = E[f(x)]$$

$$k(x, x') = E[(f(x) - m(x))(f(x') - m(x'))]$$

GPによる回帰



- シミュレーションへの応用
  - 入力と出力の関係 $f(\cdot)$ はシミュレーション・モデルの抽象的表現。
  - これまで、シミュレーションモデルは、実験や理論などからの物理的過程をプログラム・コードとして記述していくことで実装されてきた。
  - このような構成法では、(1) 定性的には現象を再現するモデルが得られても、予測性能など定量的な指標は回帰によるものに比べて劣ることがある。また、(2) 一般に膨大な計算量を必要とし、多数の設定 (f の引数 x) に対する応答を網羅的に調べるのは困難である。
  - GPによる回帰により、(1) データから直接 $f(\cdot)$ を構成する、(2) 従来の構成法によるシミュレーションの出力から軽量な $f(\cdot)$ (これをエミュレーションと呼ぶ)を構成することで、これらの問題の解決を図る。

GPによる回帰

- 平均と分散:

$$\mathbf{m} = [m(x_1), m(x_2), ..., m(x_n)]$$

$$\mathbf{K} = \begin{bmatrix} k(x_1,x_1) & k(x_1,x_2) & ... & k(x_1,x_n) \\ k(x_2,x_1) & k(x_2,x_2) & ... & k(x_2,x_n) \\ ... & ... & ... & ... \\ k(x_n,x_1) & k(x_n,x_2) & ... & k(x_n,x_n) \end{bmatrix}$$

$k(x, x')$ - any valid kernel function.

GPによる回帰

$$p(y_* \mid x_*, \mathbf{y}, \mathbf{x}) = \int \underbrace{p(y_* \mid f_*)}_{\text{Gaussian}} p(f_* \mid x_*, \mathbf{y}, \mathbf{x}) df_*$$

$$p(f_* \mid x_*, \mathbf{y}, \mathbf{x}) = \int \underbrace{p(f_* \mid \mathbf{f}, x_*, \mathbf{x})}_{\text{Gaussian}} \underbrace{p(\mathbf{f} \mid \mathbf{y}, \mathbf{x})}_{\text{Gaussian}} d\mathbf{f}$$

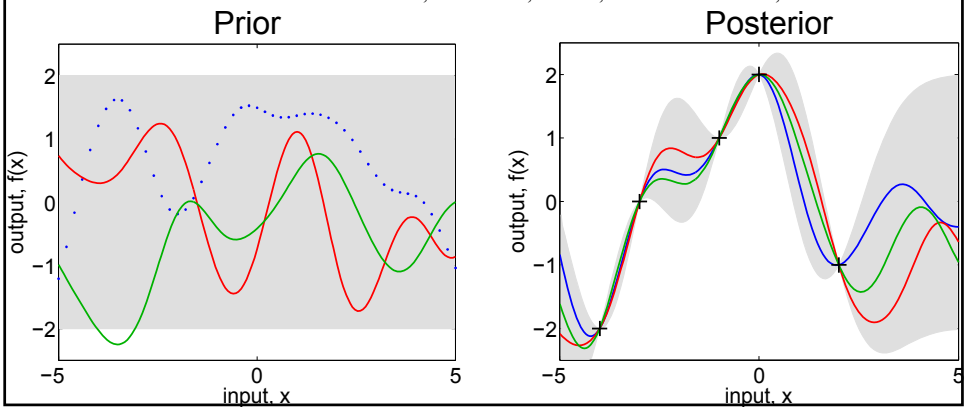
$$p(y_* \mid x_*, \mathbf{y}, \mathbf{x}) = \mathbf{N}(\bar{f}_*, \mathrm{var}(f_*))$$

GPによる回帰

- 回帰問題: $y = f(x) + \varepsilon, \quad \varepsilon \sim \mathbf{N}(0, \sigma^2)$
- 予測: $p(y_* \mid x_*, \mathbf{y}, \mathbf{x}) = \mathbf{N}(\bar{f}_*, \mathrm{var}(f_*))$

$$\bar{f}_* = \mathbf{K}_{x_*,\mathbf{x}}(\mathbf{K}_{\mathbf{x},\mathbf{x}} + \sigma^2\mathbf{I})^{-1}\mathbf{y}$$ ※ $m(x) = 0$ を仮定

$$\mathrm{var}(f_*) = \mathbf{K}_{x_*,x_*} - \mathbf{K}_{x_*,\mathbf{x}}(\mathbf{K}_{\mathbf{x},\mathbf{x}} + \sigma^2\mathbf{I})^{-1}\mathbf{K}_{\mathbf{x},x_*}$$



GPによる回帰

- 出力がベクトルの場合 ($\mathbf{f} : \mathbb{R}^p \to \mathbb{R}^q$ の推定)
  - 仮定: データ間の相関と成分間の相関が分離できる

$k(\mathbf{x}_i, \mathbf{x}_j) \times [\Sigma]_{kl} =$

i 番目の入力データに対する出力の k 成分目と

j 番目の入力データに対する出力の l 成分目の間の相関

  - 事前分布

$$\mathbf{f}(\mathbf{x}) \sim \mathcal{GP}(\mathbf{m}(\mathbf{x}), k(\mathbf{x}, \mathbf{x}')\Sigma)$$

$$\mathbf{f}(\mathbf{x}) = [f_1(\mathbf{x}), \cdots, f_q(\mathbf{x})]^T \in \mathbb{R}^q$$

$$\Sigma \in \mathbb{R}^{q \times q}$$

$$\mathbf{m}(\mathbf{x}) = E[\mathbf{f}(\mathbf{x})] \equiv \mathbf{0}$$

$$k(\mathbf{x}, \mathbf{x}')\Sigma = E[(\mathbf{f}(\mathbf{x}) - \mathbf{m}(\mathbf{x}))(\mathbf{f}(\mathbf{x}') - \mathbf{m}(\mathbf{x}'))^T]$$

(Conti & O'Hagen, Journal of Statistical Planning and Inference **140**, 640-651 (2010) の単純化)

GPによる回帰

- 出力がベクトルの場合
  - 事後分布

$$\mathbf{f}(\mathbf{x}) \mid \mathbf{X}, \mathbf{Y} \sim \mathcal{GP}(\mathbf{m}^*(\mathbf{x}), k^*(\mathbf{x},\mathbf{x}')\Sigma)$$

$$\mathbf{X} = [\mathbf{x}_1, \cdots, \mathbf{x}_n] \in \mathbb{R}^{p \times n} \text{ (訓練データ入力)}$$

$$\mathbf{Y} = [\mathbf{y}_1, \cdots, \mathbf{y}_n] \in \mathbb{R}^{q \times n} \text{ (訓練データ出力)}$$

$$\mathbf{k}(\mathbf{x}) = [k(\mathbf{x},\mathbf{x}_1), \cdots, k(\mathbf{x},\mathbf{x}_n)]^T \in \mathbb{R}^{n \times 1}$$

$$[\mathbf{K}]_{ij} = k(\mathbf{x}_i, \mathbf{x}_j); \quad \mathbf{K} \in \mathbb{R}^{n \times n}$$

$$\mathbf{m}^*(\mathbf{x}) = \mathbf{Y}[\mathbf{K} + \sigma^2 \mathbf{I}_{n \times n}]^{-1}\mathbf{k}(\mathbf{x})$$

$$k^*(\mathbf{x},\mathbf{x}') = k(\mathbf{x},\mathbf{x}') - \mathbf{k}^T(\mathbf{x})\mathbf{K}^{-1}\mathbf{k}(\mathbf{x}')$$

  - 今回の話では、$\sum$ を対角に取るので、独立なスカラーのGPを成分の数だけ使ったのと等価。

GPによる回帰

- ハイパーパラメータの学習：

$$\Theta = \{\sigma, l\}$$

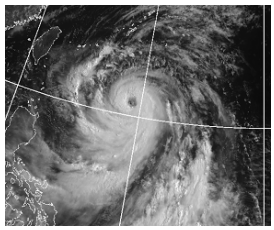
$$\max_{\Theta} p(\mathbf{y} \mid \mathbf{x}, \Theta) = \max_{\Theta} \int p(\mathbf{y} \mid \mathbf{f}) p(\mathbf{f} \mid \mathbf{x}, \Theta) d\mathbf{f}$$

ML type –II estimation

  - 今回の話では、この計算に粒子群最適化を使う

台風

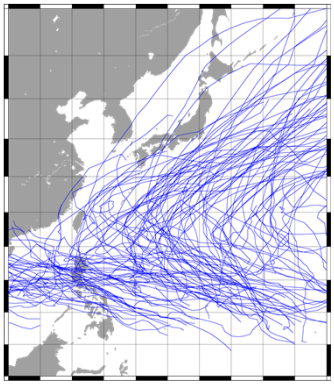
- 北西太平洋域で発生する強い熱帯低気圧.
- 東アジア域においては，深刻な災害をしばしば引き起こす.
- したがって，台風による災害のリスクを評価することが重要.
- 多様なシナリオを考慮することで，確率的なリスク評価につなげたい.
- 特に，地球温暖化等による気候変動の効果を考慮した評価が，長期的には重要.



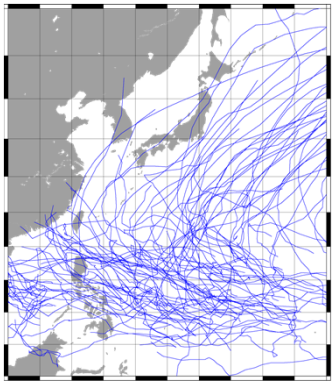
(c) JMA

(c) BBC

確率台風モデル



August from 1980 to 2009

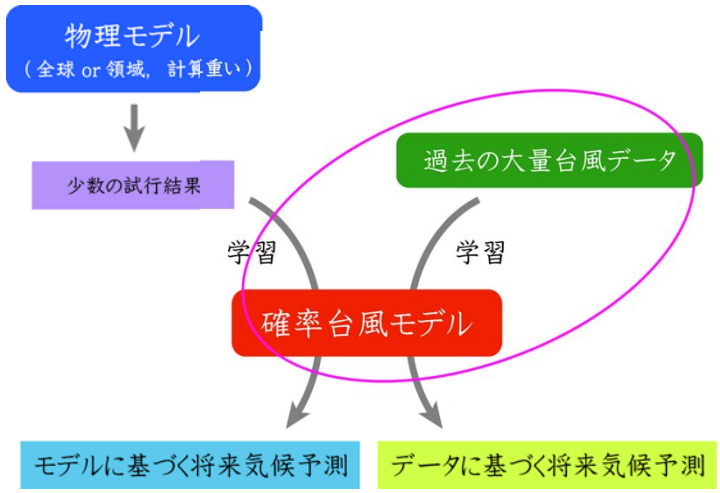


Generated from the model

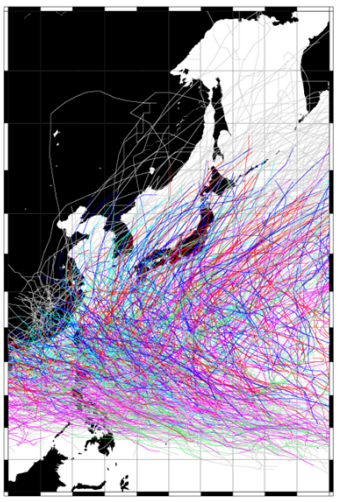
台風による災害リスクの見積もりのため，しばしば，台風の挙動を表現する確率モデルが用いられる.

過去のデータから，起こりそうなシナリオを色々と提示するのが目的.

将来気候予測と確率モデル



Best track data



“Pollock plot” of the best track data

- 気象庁(The Regional Specialized Meteorological Center)のbest track dataを使用.
- 1951年以来の60年分以上の台風について，位置，中心気圧，中心風速などが6時間ごとに記録されている.
- 今回は位置のデータだけを使用.

確率台風モデル

- 台風の移動速度は，背景の風でかなり決まる.
- そこで，個々の台風の速度を以下のように与えることが多い:

$$\boldsymbol{v}_t = \boldsymbol{V}(\phi_t, \lambda_t) + \boldsymbol{v}_{res,t},$$

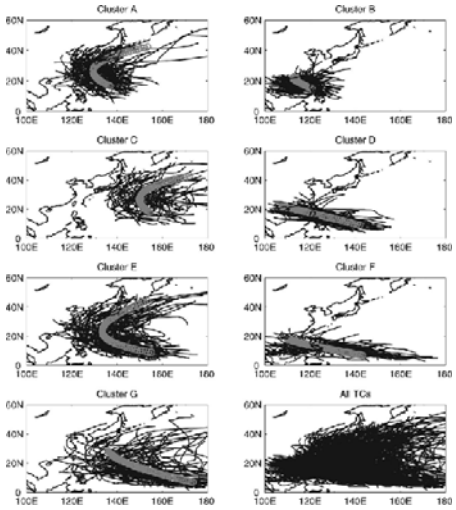
ただし，$\boldsymbol{V}$ は背景の(平均的な)台風移動速度場である.

- $\boldsymbol{v}_{\mathrm{res},t}$ は以下のようにARモデルを使って与えられることが多い(e.g., Vickery et al., 2000):

$$\boldsymbol{v}_{\mathrm{res},t} = \alpha \boldsymbol{v}_{\mathrm{res},t-1} + \boldsymbol{e}_t, \quad (\boldsymbol{e}_t : \text{Gaussian perturbation}).$$

- 本研究では，$\boldsymbol{V}$ の設定について検討.

これまでの研究



(Camargo et al., 2007)

平均的な風速分布だけで多様な軌道を生成するのは難しい.

軌道パターンや擾乱成分などの，クラスタリングによる場合分けがよく用いられる.

- いずれかのクラスタからのサンプリングで台風を合成する

多様な風速分布を表現できるようなモデルを作れないか.

移動速度モデル

- 台風の移動速度 $\boldsymbol{v}$ が，緯度，経度，年内日付，年の4変数の関数で表現できると仮定し，入力に対する出力vを予測するモデルをガウス過程回帰で生成 (e.g., Kennedy and O'Hagan, 2000).

$$\boldsymbol{v}(z) \sim \mathcal{GP}\left[\begin{pmatrix}0\\0\end{pmatrix}, k(z,z')\begin{pmatrix}\alpha_\phi & 0\\ 0 & \alpha_\lambda\end{pmatrix}\right], \quad \boldsymbol{\alpha} = \begin{pmatrix}\alpha_\phi\\ \alpha_\lambda\end{pmatrix}$$

$$z = (\phi \;\; \lambda \;\; d \;\; Y) \quad (\phi: \text{経度}; \lambda: \text{緯度}; d: \text{年内日付}; Y: \text{年}).$$

- 共分散関数 $k(z,z')$ は，

$$k(z,z') = \exp\left[-\frac{(\phi-\phi')^2}{\sigma_\phi^2} - \frac{(\lambda-\lambda')^2}{\sigma_\lambda^2} - \frac{(d-d')^2}{\sigma_d^2} - \frac{(Y-Y')^2}{\sigma_Y^2}\right]$$

と与える.

移動速度場の推定

- データ生成: Best track dataの台風位置情報 $\{\boldsymbol{x}(z_{n,i})\}_{n,i}$ (n: 台風の識別子; i: 時間ステップ) から，台風の速度を

$$\boldsymbol{v}(z_{n,i}) = \left(\boldsymbol{x}(z_{n,i+1}) - \boldsymbol{x}(z_{n,i})\right) / \text{6時間}$$

で求める.

  - データ総数 = 数日/6時間 ×30個/年 × 30年 ≈ 30,000

- 観測モデル: GPの出力 + 正規ノイズ = 観測される台風の速度

$$\boldsymbol{v}^{\text{obs}}(z) \quad \sim \quad N(\boldsymbol{v}(z), \sigma_0^2 I)$$

移動速度場の推定

- 潜在変数と観測データの関係:

$$\boldsymbol{v}(z) \sim \mathcal{GP}\left(\mathbf{0}, k(z,z')\,\mathrm{diag}\,\boldsymbol{\alpha}\right)$$

$$\boldsymbol{v}^{\mathrm{obs}}(z) \sim N(\boldsymbol{v}(z), \sigma_0^2 I)$$

- データ $\mathsf{V} = \left(\boldsymbol{v}_1^{\mathrm{obs}} \cdots \boldsymbol{v}_N^{\mathrm{obs}}\right)$ から得られる事後分布の平均を推定値として用いる。正規分布に関する公式から、この推定値は

$$\hat{\boldsymbol{v}}(z_i) \equiv E[\boldsymbol{v}(z_i) \mid \mathsf{V}] = \boldsymbol{k}(z_i)(\mathsf{K} + \sigma_0^2 \mathsf{I})^{-1}\mathsf{V}$$

となる. 但し,

$$\boldsymbol{k}(z) = \left(k(z, z_{1,1}) \;\cdots\; k(z, z_{N,T_N})\right)^T, \quad K = \begin{bmatrix} k(z_{1,1}, z_{1,1}) & \cdots & k(z_{1,1}, z_{N,T_N}) \\ \vdots & \ddots & \vdots \\ k(z_{N,T_N}, z_{1,1}) & \cdots & k(z_{N,T_N}, z_{N,T_N}) \end{bmatrix}.$$

移動速度場の推定

- $k(z,z')$ に含まれるパラメータは5分割クロスバリデーションで決定.
  - データ・グループの分割 (5年おきにグルーピング)

$$\mathsf{Z}_1 = \{z \mid Y(z) \in \{1980, 1985, 1990, 1995, 2000, 2005\}\}$$
$$\mathsf{Z}_2 = \{z \mid Y(z) \in \{1981, 1986, 1991, 1996, 2001, 2006\}\}$$
$$\mathsf{Z}_3 = \{z \mid Y(z) \in \{1982, 1987, 1992, 1997, 2002, 2007\}\}$$
$$\mathsf{Z}_4 = \{z \mid Y(z) \in \{1983, 1988, 1993, 1998, 2003, 2008\}\}$$
$$\mathsf{Z}_5 = \{z \mid Y(z) \in \{1984, 1989, 1994, 1999, 2004, 2009\}\}$$

  - パラメータ $\boldsymbol{\sigma} = (\sigma_\lambda, \sigma_\phi, \sigma_d, \sigma_Y; \sigma_0)$ に対するエラー関数

$$E_g(\boldsymbol{\sigma}) = \sum_{z \in Z_g} \left(E\left[\boldsymbol{v}(z) \middle| \left\{ \boldsymbol{v}^{\mathrm{obs}}(z) \middle| z \in \bigcup_{g' \neq g} \mathsf{Z}_{g'} \right\}, \boldsymbol{\sigma} \right] - \boldsymbol{v}^{\mathrm{obs}}(z) \right)^2; \quad E(\boldsymbol{\sigma}) = \sum_g E_g(\boldsymbol{\sigma})$$

  - 粒子群最適化により、$\boldsymbol{\sigma}$ を決定する。

移動速度場の推定

- 粒子群最適化: 魚や昆虫群の索餌行動を模倣. 1匹(粒子)が餌を見つけると, 他の虫がそちらへ移動方向を変更.
  - パラメータ: w (慣性定数; 1より少し小さい), c_1, c_2 (良い位置に向かう程度; 1程度), $E(x)$ (適応度関数)
  - 初期化: $x_0^{(i)}$, $v_0^{(i)}$ を適当に設定して $E_0^{(i)} = E(x_0^{(i)})$, $i = 1, \ldots, I$.
  - for $n = 1, \cdots, N$ (n : ステップ数)

    for $i = 1, \cdots, I$ (i : 粒子番号)

    $x_n^{(i)} = x_{n-1}^{(i)} + v_n^{(i)}$; $E_n^{(i)} = E(x_n^{(i)})$ (移動・スコア計算)

    $(\hat{x}_n^{(i)}, \hat{E}_n^{(i)}) = \mathrm{Min}\,\{(\hat{x}_{n-1}^{(i)}, \hat{E}_{n-1}^{(i)}); (x_n^{(i)}, E_n^{(i)})\}$ (ローカル・ベスト)

    $(\hat{x}_n, \hat{E}_n) = \mathrm{Min}\,\{(\hat{x}_{n-1}, \hat{E}_{n-1})\} \cup \{(\hat{x}_n^{(i)}, \hat{E}_n^{(i)})\}_{i=1}^{N}$ (グローバル・ベスト)

    for $i = 1, \cdots, I$

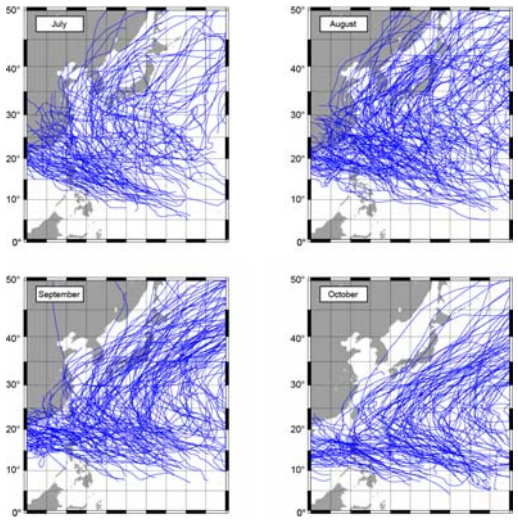
    $v_n^{(i)} = w v_{n-1}^{(i)} + r_1(\hat{x}_n^{(i)} - x_n^{(i)}) + r_2(\hat{x}_n - x_n^{(i)})$ (速度更新)

    where $r_1 \sim U[0, c_1)$; $r_2 \sim U[0, c_2)$

移動速度場の推定

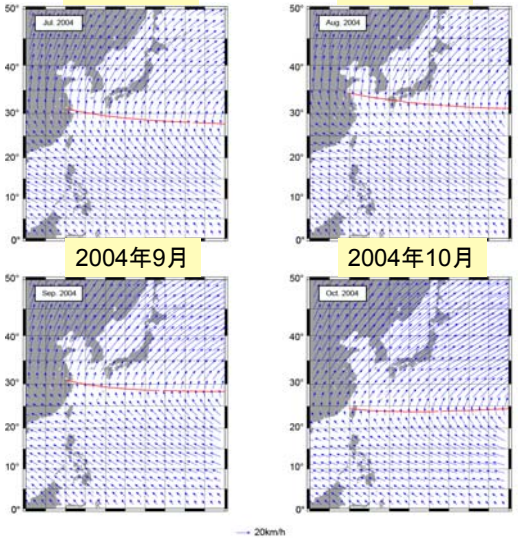
- 推定値
  - $\sigma_\lambda \approx 0.13^\circ$, $\sigma_\phi \approx 0.66^\circ$, $\sigma_d \approx 30\,\mathrm{days}$, $\sigma_Y \approx 200\,\mathrm{years}$
- 6時間ごとの台風移動速度 $\boldsymbol{v}$ が独立に GP から生成されているかのように扱っているので, ランダムに分割すると, σ_d, σ_Y が極めて小さく見積もられてしまう. (前後の時間ステップの $\boldsymbol{v}$ を補間すると, かなりうまく推定できてしまう.)

Best track data (1980-2013)



季節変動

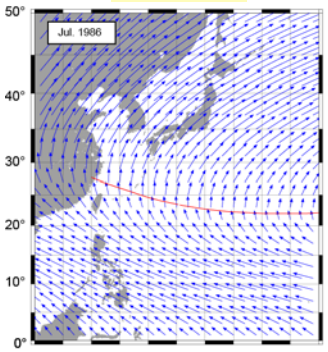
2004年7月　2004年8月



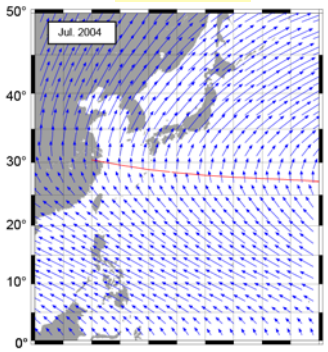
2004年9月　2004年10月

長期的な変動

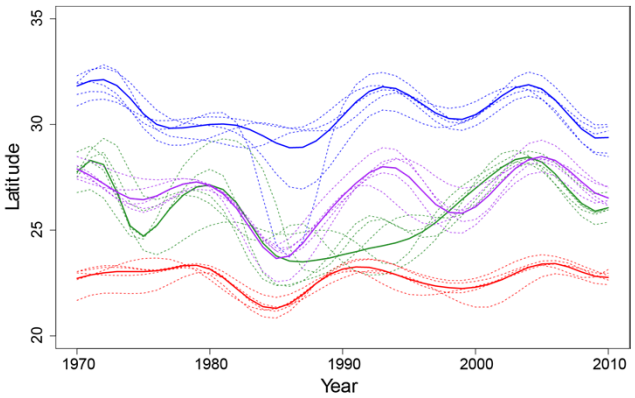
1986年7月



2004年7月

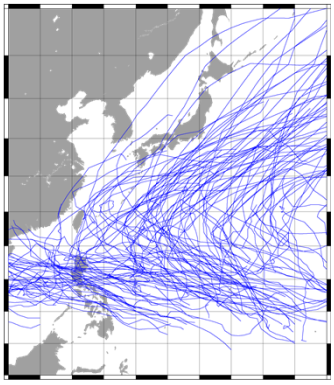


長期的な変動

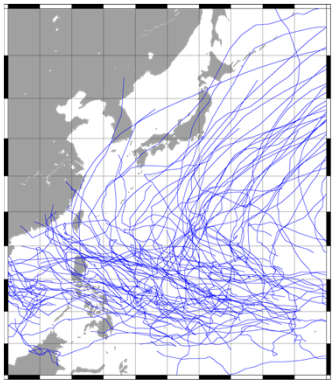


$\sigma_\lambda \approx 0.13^\circ, \quad \sigma_\phi \approx 0.66^\circ, \quad \sigma_d \approx 30\,\text{days}, \quad \sigma_Y \approx 4\,\text{years}$

結果の例



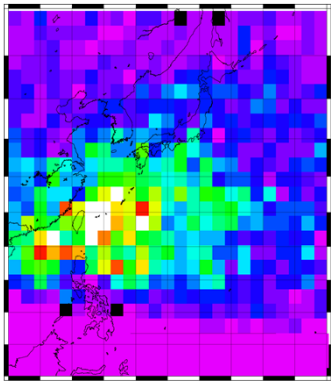
August from 1980 to 2009



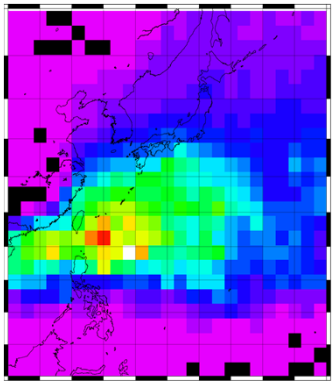
Generated from the model

それらしい軌道は出るようになった.

結果の例



August from 1980 to 2009



Estimate by the model

実は，モデルでは倍くらい台風を生成しないと頻度分布が合わない.

Summary

- 台風の移動速度の時空間分布のモデル化を試みた.
- 移動速度分布パターンの季節変動の特徴は, うまく表現できるようになった.
- 長期変動の再現についてはあやしい.
  - σ_Y = 200年 >> T = 30年なので、データからトレンドの存在は主張できない。
- 頻度分布の改善が今後の課題.
  - 生成・消滅機構のモデルの改善.
  - 擾乱成分のモデルの改善.