

# Modelos predictivos aplicados a datos climáticos y dengue: un enfoque espacio-temporal y de aprendizaje de máquinas.

Shu Wei Chou-Chen

Escuela de Estadística - Centro de Investigación en Matemática Pura y Aplicada (CIMPA)

Universidad de Costa Rica

25 de mayo, 2023



UNIVERSIDAD DE  
**COSTA RICA**

EEs

Escuela de  
Estadística

**CIMPA**

Centro de Investigación en  
Matemática Pura y Aplicada

# Contenidos

- 1 Introducción**
- 2 Modelos predictivos de riesgos de dengue con variables climáticas**
- 3 Emulador espacio-temporal climático (downscaling)**
- 4 Conclusiones**

# Introducción

- El dengue es una enfermedad sensible al clima (temperatura, humedad, precipitación, etc.).
- Afecta la biología, el comportamiento y la disponibilidad del mosquito para reproducirse, desarrollarse, propagar el virus e interactuar con el huésped humano.
- El uso de imágenes satelitales y monitoreo del clima como datos de entrada en modelos de aprendizaje automático y otros enfoques de aprendizaje estadístico han mostrado resultados prometedores que podrían predecir de manera efectiva el riesgo relativo de transmisión del dengue.



# Contenido

## 1 Introducción

## 2 Modelos predictivos de riesgos de dengue con variables climáticas

- Datos
- Modelo para cada cantón
  - Entrenamiento del modelo
  - Predicción
  - Resultados
- Modelo espacio-temporal
  - Metodología
  - Resultados

## 3 Emulador espacio-temporal climático (downscaling)

- Métodos de aproximación

## 4 Conclusiones

# Preliminares

En Costa Rica:

- Circulación endémica de tres de los cuatro serotipos del virus del dengue (DENV-1, DENV-2, DENV-3).
- Clima tropical, que proporciona condiciones ideales para que el vector del mosquito sobreviva, se replique y transmita la enfermedad.
- Microclimas separados por distancias cortas hacen que sea crucial personalizar el análisis de riesgo de transmisión del dengue en un país.



# Preliminares

- Las autoridades sanitarias no utilizan formalmente la información meteorológica como entrada para desarrollar actividades de prevención y control.
- Una colaboración inicial con las autoridades sanitarias costarricenses permitió identificar 32 municipios (cantones) de interés basados en sus características entomológicas y epidemiológicas.

# Preliminares

- Se utilizaron dos enfoques de modelado diferentes para predecir el riesgo relativo de infecciones por dengue en 32 cantones en Costa Rica:
  - Barboza LA, Chou-Chen SW, Vásquez P, García YE, Calvo JG, et al. (2023) **Assessing dengue fever risk in Costa Rica by using climate variables and machine learning techniques**. PLOS Neglected Tropical Diseases 17(1): e0011047.  
<https://doi.org/10.1371/journal.pntd.0011047>
  - Chou-Chen, S. W., Barboza, L. A., Vásquez, P., García, Y. E., Calvo, J. G., Hidalgo, H. G., & Sanchez, F. (2023). **Bayesian spatio-temporal model with INLA for dengue fever risk prediction in Costa Rica**. arXiv preprint arXiv:2302.06747.

# Contenidos

## 1 Introducción

## 2 Modelos predictivos de riesgos de dengue con variables climáticas

- Datos
- Modelo para cada cantón
  - Entrenamiento del modelo
  - Predicción
  - Resultados
- Modelo espacio-temporal
  - Metodología
  - Resultados

## 3 Emulador espacio-temporal climático (downscaling)

- Métodos de aproximación

## 4 Conclusiones

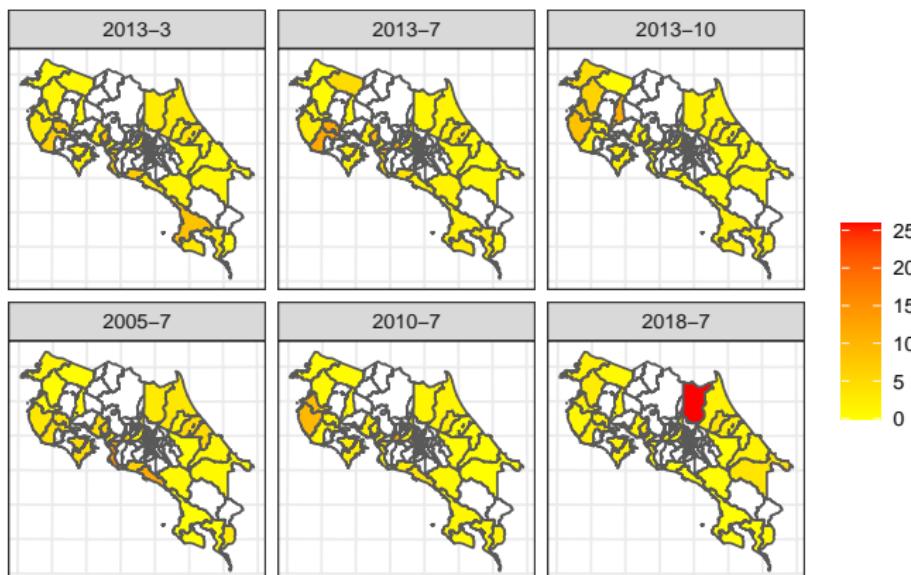
# Datos

## Dengue:

- Casos mensuales sospechosos y confirmados clínicamente de fiebre del dengue en Costa Rica desde el año 2000 hasta 2021.
- Riesgo relativo:

$$RR_{i,t} = \frac{\frac{\text{Casos}_{i,t}}{\text{Población}_{i,t}}}{\frac{\text{Casos}_{CR,t}}{\text{Población}_{CR,t}}}$$

# Datos



**Figura 1:** Riesgo Relativo ( $RR$ ) de los 32 cantones en estudio para diferentes meses y años de datos disponibles.

# Datos

## Clima:

- **Cálculos de precipitación diaria ( $P_{i,t}$ )**: obtenido de datos de Climate Hazards Group InfraRed Precipitation with Station data (CHIRPS).
- **Anomalía de la temperatura de la superficie del mar ENSO ( $S_{i,t}$ )**: obtenido del Climate Prediction Center (CPC), National Oceanographic and Atmospheric Administration (NOAA).
- **Índice de Vegetación de Diferencia Normalizada (NDVI) ( $N_{i,t}$ )**: obtenido del Moderate Resolution Imaging Spectroradiometer (MODIS).
- **Temperatura de la superficie terrestre durante el día ( $L_{i,t}$ )**: obtenida de MODIS.
- **Índice Tropical del Atlántico Norte ( $TN_{i,t}$ )**: obtenido de NOAA.

# Contenidos

## 1 Introducción

## 2 Modelos predictivos de riesgos de dengue con variables climáticas

- Datos
- **Modelo para cada cantón**
  - Entrenamiento del modelo
  - Predicción
  - Resultados
- **Modelo espacio-temporal**
  - Metodología
  - Resultados

## 3 Emulador espacio-temporal climático (downscaling)

- Métodos de aproximación

## 4 Conclusiones

# Entrenamiento del modelo

- Para cada cantón fijo  $i$ ,  $i = 1, \dots, 32$ , dividimos el conjunto de datos en:

**Período de calibración:** de enero de 2000 a diciembre de 2020.

**Período de prueba:** de enero de 2021 a marzo de 2021.

- Incorporación de las asociaciones históricas retrasadas de las covariables climáticas aplicando un marco de modelo no lineal de retardo distribuido (DLNM) (Gasparrini, 2014; Gasparrini et al., 2010).
- Consiste en un espacio bidimensional de funciones que especifica una función de exposición-retardo-respuesta, que depende del predictor  $x$  a lo largo de los retardos temporales  $\ell$ :

$$s(x; t) = \int_{\ell_0}^{\ell_1} f \cdot w(x_{x-\ell}, \ell) d\ell \approx \sum_{\ell=\ell_0}^{\ell_1} f \cdot w(x_{x-\ell}, \ell). \quad (1)$$

# Entrenamiento del modelo

La estructura de los modelos en términos de la variable dependiente y las covariables para un cantón  $i$  es la siguiente:

$$RR_t \sim f(RR_{t-1}, C_1 P_t, C_2 S_t, C_3 N_t, C_4 L_t, C_5 TN_t, M_t) \quad (2)$$

donde

$f$  es una función que depende del método (GAMLSS o RF),  
las matrices  $C_i$  se definen en términos de la representación DLNM, y  
 $M_t$  es una variable de tipo factor que describe el efecto fijo mensual.

# Entrenamiento del modelo

## 1 GAMLSS:

$$RR_t \stackrel{ind}{\sim} \mathcal{D}(\mu, \sigma, \nu)$$

$$\begin{aligned} g_1(\mu) = & \beta_{10} + \beta_{11} RR_{t-1} + \beta_{12} C_1 P_t + \beta_{13} C_2 S_t \\ & + \beta_{14} C_3 N_t + \beta_{15} C_4 L_t + \beta_{16} C_5 TN_t + \beta_{17} M_t \end{aligned}$$

$$g_2(\sigma) = \beta_{20}$$

$$g_3(\nu) = \beta_{30}$$

donde  $\mathcal{D}$  es la distribución gamma ajustada a cero (ZAGA):

$$f_Y(y) = \begin{cases} \nu & \text{si } y = 0; \\ (1 - \nu)f_W(y) & \text{si } 0 < y < \infty. \end{cases}$$

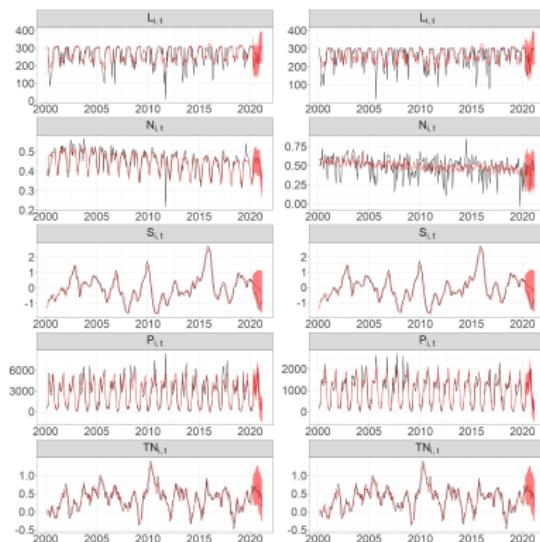
para  $0 \leq y < \infty$ , donde  $W \sim GA(\mu, \sigma)$  es la distribución gamma con  $0 < \mu < \infty$ ,  $0 < \sigma < \infty$  y  $0 < \nu < 1$

## 2 Bosque aleatorio: Un método de conjunto que consiste en una gran cantidad de árboles de decisión.

# Predicción

- Se utiliza el modelo vectorial autorregresivo (VAR) para cada cantón para predecir las covariables climáticas en el período de prueba, ya que las covariables climáticas utilizadas en este estudio están altamente correlacionadas.
- Estas predicciones climáticas, junto con los riesgos relativos predichos, se utilizan para proporcionar pronósticos de la variable dependiente durante el período de prueba.

# Predicción



**Figura 2:** Covariables climáticas observadas y valores de pronóstico en dos cantones específicos: Alajuela (paneles izquierdos) y Quepos (paneles derechos). Línea negra: covariables climáticas observadas, línea roja: valores de pronóstico y áreas sombreadas en rojo: regiones de confianza del 95 %.

# Predicción

- Aplicamos bootstrap no paramétrico para construir intervalos de predicción.

## Comparación de modelos:

- 1 Raíz de error cuadrático medio normalizado (*NRMSE*):

$$NRMSE = \sqrt{\frac{1}{m\overline{RR}} \sum_{t=1}^m (RR_t - \widehat{RR}_t)^2}$$

- 2 Puntuación de intervalo (interval score) normalizado (*NIS*<sub>α</sub>) con  $\alpha$ :

$$\begin{aligned} NIS_\alpha = \frac{1}{m\overline{RR}} \sum_{t=1}^m & \left[ (U_t - L_t) + \frac{2}{1-\alpha} (L_t - RR_t) \cdot 1_{RR_t < L_t} \right. \\ & \left. + \frac{2}{1-\alpha} (RR_t - U_t) \cdot 1_{RR_t > U_t} \right], \end{aligned}$$

# Resultados

**Cuadro 1:** Especificación DLNM para covariables climáticas (máximo de 18 meses)

| Covariable  | esp. variable | esp. rezago |
|---|---------------|-------------|
| Precipitación ( $P_{i,t}$ )   | B-spline      | lineal      |
| Anomalía de la temperatura de la superficie del mar ( $S_{i,t}$ )   | B-spline      | lineal      |
| Índice de vegetación de diferencia normalizada (NDVI) ( $N_{i,t}$ ) | lineal        | lineal      |
| Temperatura de la superficie terrestre diurna ( $L_{i,t}$ )         | lineal        | lineal      |
| Índice tropical del Atlántico norte ( $TN_{i,t}$ )                  | lineal        | lineal      |

- Estas elecciones permiten un equilibrio aceptable entre la complejidad de los modelos y la precisión predictiva en todas las ubicaciones.
- Por lo tanto, podemos predecir el riesgo relativo de dengue para los primeros tres meses de 2021.

# Resultados

Cuadro 2: Mejor modelo para cada cantón

| Cantón        | NRMSE   | NIS <sub>95</sub> | Mejor Modelo |
|---------------|---------|-------------------|--------------|
| Alajuela      | 0.2218  | 2.2789            | RF           |
| Alajuelita    | 0.2650  | 13.9639           | GAMLSS       |
| Atenas        | 1.7615  | 16.6668           | GAMLSS       |
| Cañas         | 0.4572  | 8.8197            | GAMLSS       |
| Carrillo      | 0.7468  | 13.2383           | GAMLSS       |
| Corredores    | 7.4754  | 32.9063           | RF           |
| Desamparados  | 0.0333  | 0.9103            | RF           |
| Esparza       | 0.5269  | 13.5993           | RF           |
| Garabito      | 32.7929 | 151.7930          | GAMLSS       |
| Golfito       | 0.7657  | 13.0531           | RF           |
| Guácimo       | 1.9182  | 8.1629            | RF           |
| La Cruz       | 9.1306  | 127.2412          | RF           |
| Libería       | 2.2955  | 57.4666           | RF           |
| Limón         | 2.1313  | 18.4764           | RF           |
| Matina        | 3.9168  | 31.3834           | GAMLSS       |
| Montes de Oro | 18.9354 | 162.9258          | RF           |
| Nicoya        | 5.3693  | 47.4291           | GAMLSS       |

# Resultados

| Cantón        | NRMSE   | NIS <sub>95</sub> | Mejor modelo |
|---------------|---------|-------------------|--------------|
| Orotina       | 0.0860  | 0.6048            | RF           |
| Osa           | 2.4703  | 2.8810            | GAMLSS       |
| Parrita       | 8.0959  | 33.1650           | GAMLSS       |
| Perez Zeledón | 1.7181  | 14.7112           | GAMLSS       |
| Pococí        | 0.6451  | 10.2410           | RF           |
| Puntarenas    | 1.7821  | 22.3798           | GAMLSS       |
| Quepos        | 34.3754 | 179.4324          | GAMLSS       |
| San José      | 0.0195  | 2.5938            | RF           |
| Santa Ana     | 0.6331  | 31.8002           | RF           |
| Santa Cruz    | 0.7159  | 13.4637           | GAMLSS       |
| Sarapiquí     | 0.5572  | 1.7992            | RF           |
| Siquirres     | 0.8202  | 6.6540            | RF           |
| Talamanca     | 10.9225 | 26.2873           | RF           |
| Turrialba     | 0.9400  | 7.0685            | GAMLSS       |

# Resultados

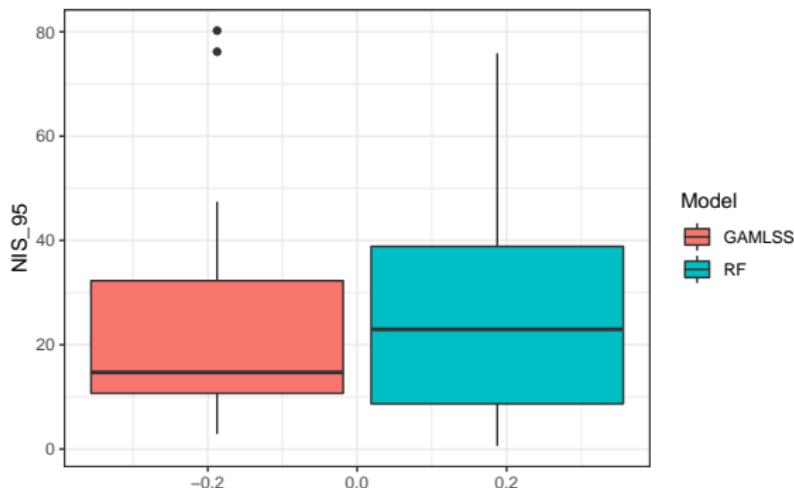
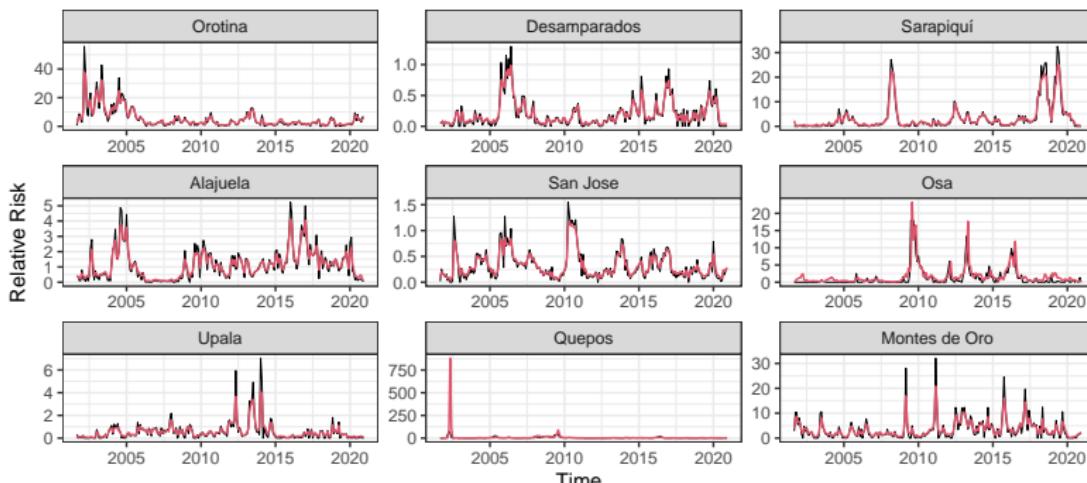


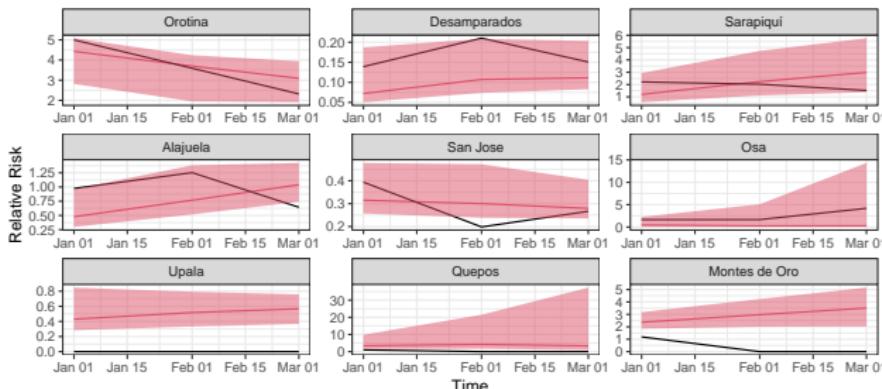
Figura 3: Comparación de la distribución de la métrica NIS entre los métodos.

# Resultados



**Figura 4:** Comparación durante el periodo de ajuste. Los seis paneles superiores muestran los mejores cantones según la métrica NIS. Los tres paneles inferiores muestran los peores cantones según la métrica NIS. La línea negra representa el RR observado, la línea roja representa el RR estimado.

# Resultados



**Figura 5:** Comparación de pronósticos durante el periodo de prueba (2021). Los seis paneles superiores muestran los mejores cantones según la métrica NIS. Los tres paneles inferiores muestran los peores cantones según la métrica NIS.

# Contenidos

## 1 Introducción

## 2 Modelos predictivos de riesgos de dengue con variables climáticas

- Datos
- Modelo para cada cantón
  - Entrenamiento del modelo
  - Predicción
  - Resultados
- **Modelo espacio-temporal**
  - Metodología
  - Resultados

## 3 Emulador espacio-temporal climático (downscaling)

- Métodos de aproximación

## 4 Conclusiones

# Metodología

Para cada cantón  $i$  y el tiempo  $t$ :

$$Y_{it} | \mu_{it}, \kappa \sim NegBin(\mu_{it}, \kappa)$$

donde

$$\log(\mu_{it}) = \log(E_{it}) + \log(RR_{it})$$

$$\begin{aligned} \log RR_{it} = & \alpha + f_1(RR_t) + f_2(P_t) + f_3(S_t) \\ & + f_4(N_t) + f_5(L_t) + f_6(TN_t) + f_7(M_t) \\ & + \phi_{i,(month)} + \theta_{i,(year)}, \end{aligned}$$

$$\phi_{i,(month)} - \phi_{i,(month-1)} \sim N(0, \sigma_\phi^2), \text{ y}$$

$f_k, k = 1, \dots, 7$  es la estructura de DLNM aplicada desde el rezago 3 hasta el 12.

# Metodología

- Para el efecto espacial, se definen dos tipos de matriz de proximidad  $\mathbf{W}$ :
  - ① **La matriz de vecinos** usual se define como  $\mathbf{W} = \{\mathbf{W}\}_{ij} = 1$  si los municipios  $i$  y  $j$  son vecinos y 0 en caso contrario.
  - ② **Una matriz de distancia alternativa** basada en la distancia de la carretera principal en kilómetros entre el centro de cada par de municipios, es decir,  $\mathbf{W} = \{\mathbf{W}\}_{ij} = 1$  si la distancia es menor que la mediana general y 0 en caso contrario. Incorporamos esta distancia para proporcionar una forma más realista de medir la proximidad entre las dinámicas sociales.

# Metodología

- Para el efecto espacial estructurado, se define el modelo autorregresivo condicional instrínseco (CAR):

$$\theta_{i,(year)} | \theta_{j,(year)}, \tau_\theta \sim N \left( \frac{1}{n_i} \sum_{j \sim i} \theta_{j,(year)}, \frac{1}{\tau_\theta n_i} \right),$$

donde  $\tau_\theta$  es la precisión condicional,  $j \sim i$  denota que  $W_{ij} = 1$ , y  $n_i$  es el número de vecinos según la definición de la matriz de proximidad.

- En resumen, 4 estructuras espaciales fueron implementadas:
  - Independencia.
  - CAR.
  - CAR propia: agregando un valor positivo  $d$  a  $n_i$ .
  - Besag-York-Mollie (BYM): incluye el efecto aleatorio no estructurado.

# Resultados

**Cuadro 5:** Comparación de los modelos según el criterio de información de devianza (DIC) y log score de validación cruzada (CV log-score)

| DLNM       | Matriz de proximidad | Estructura espacial | DIC             | CV log-score  |
|------------|----------------------|---------------------|-----------------|---------------|
| Lineal*    | Vecino               | Independiente       | 57135.37        | 3.8710        |
|            |                      | CAR                 | 54256.47        | 3.6872        |
|            |                      | proper CAR          | <b>52628.40</b> | <b>3.5774</b> |
|            | Distancia            | BYM                 | 52632.24        | 3.5784        |
|            |                      | CAR                 | 53416.92        | 3.6264        |
|            |                      | proper CAR          | 52633.29        | 3.5787        |
| No lineal* | Vecino               | BYM                 | 52636.92        | 3.5787        |
|            |                      | Independiente       | 53429.63        | 3.8756        |
|            |                      | CAR                 | 50640.66        | 3.6838        |
|            | Distancia            | proper CAR          | <b>49438.81</b> | <b>3.5954</b> |
|            |                      | BYM                 | 49461.01        | 3.5977        |
|            |                      | CAR                 | 54674.34        | 3.9653        |
|            | Distancia            | proper CAR          | 49468.89        | 3.5985        |
|            |                      | BYM                 | 49456.27        | 3.5971        |

\* El mejor modelo para cada especificación DLNM está marcado en negrita.

# Resultados

Cuadro 6: Métricas predictivas del conjunto de datos de entrenamiento y prueba del modelo seleccionado.

| Cantón        | Datos de entrenamiento |                   |            |                     | Datos de prueba |                     |            |                     |
|---------------|------------------------|-------------------|------------|---------------------|-----------------|---------------------|------------|---------------------|
|               | Independiente          |                   | CAR propia |                     | Independiente   |                     | CAR propia |                     |
|               | NRMSE                  | NIS <sub>95</sub> | NRMSE      | NIS <sub>0,05</sub> | NRMSE           | NIS <sub>0,05</sub> | NRMSE      | NIS <sub>0,05</sub> |
| Alajuela      | 0.7815                 | 14.5297           | 0.3982     | 5.3710              | 0.0750          | 1.0905              | 0.0416     | 1.6417              |
| Alajuelita    | 0.3090                 | 23.2238           | 0.2175     | 13.1177             | 0.4515          | 22.8135             | 0.0515     | 2.0922              |
| Atenas        | 4.6100                 | 24.7736           | 2.5706     | 10.1453             | 5.1453          | 87.6173             | 0.3283     | 10.2199             |
| Cañas         | 10.1008                | 24.3243           | 5.7069     | 12.3277             | 5.8803          | 60.7800             | 0.4829     | 6.8129              |
| Carrillo      | 4.3786                 | 15.9598           | 3.5404     | 9.1421              | 0.9860          | 13.2041             | 0.4175     | 2.5945              |
| Corredores    | 5.0620                 | 25.1830           | 2.5155     | 8.6126              | 1.6824          | 17.0115             | 0.5697     | 1.9358              |
| Desamparados  | 0.2351                 | 18.6300           | 0.1466     | 8.2922              | 0.0282          | 3.0978              | 0.0142     | 3.0085              |
| Esparza       | 4.7287                 | 17.4775           | 2.6296     | 8.4081              | 1.1899          | 16.1291             | 0.1849     | 2.1748              |
| Garabito      | 38.4940                | 29.0191           | 5.2545     | 9.8071              | 28.8475         | 232.7807            | 0.4258     | 12.9528             |
| Golfito       | 3.7245                 | 23.5563           | 1.7734     | 8.2174              | 2.0078          | 35.7736             | 0.3499     | 8.5601              |
| Guacimo       | 1.9057                 | 13.8120           | 1.0844     | 4.4132              | 3.0709          | 18.1805             | 2.0633     | 4.8628              |
| La Cruz       | 6.2484                 | 26.6755           | 4.2779     | 14.3090             | 6.0351          | 118.6721            | 0.5029     | 14.7439             |
| Liberia       | 3.6829                 | 23.1675           | 2.0260     | 10.2626             | 4.0913          | 92.4202             | 0.5693     | 16.2507             |
| Limon         | 2.9640                 | 17.1259           | 1.6025     | 6.7593              | 1.3724          | 11.2322             | 0.1034     | 1.7042              |
| Matina        | 5.1143                 | 19.9086           | 2.7162     | 7.4890              | 58.0750         | 192.2903            | 0.1232     | 3.0869              |
| Montes de Oro | 5.9220                 | 21.8260           | 4.1289     | 12.2332             | 11.4909         | 126.5836            | 1.2106     | 19.5139             |
| Nicoya        | 2.8036                 | 20.5059           | 2.0672     | 10.1305             | 3.6262          | 101.6524            | 0.1894     | 10.2497             |
| Orotina       | 13.7161                | 17.5376           | 4.0681     | 6.0472              | 1.0599          | 6.7613              | 0.4888     | 1.5221              |
| Osa           | 5.9295                 | 33.7053           | 4.2208     | 17.2559             | 1.1235          | 13.0806             | 0.5758     | 1.7891              |

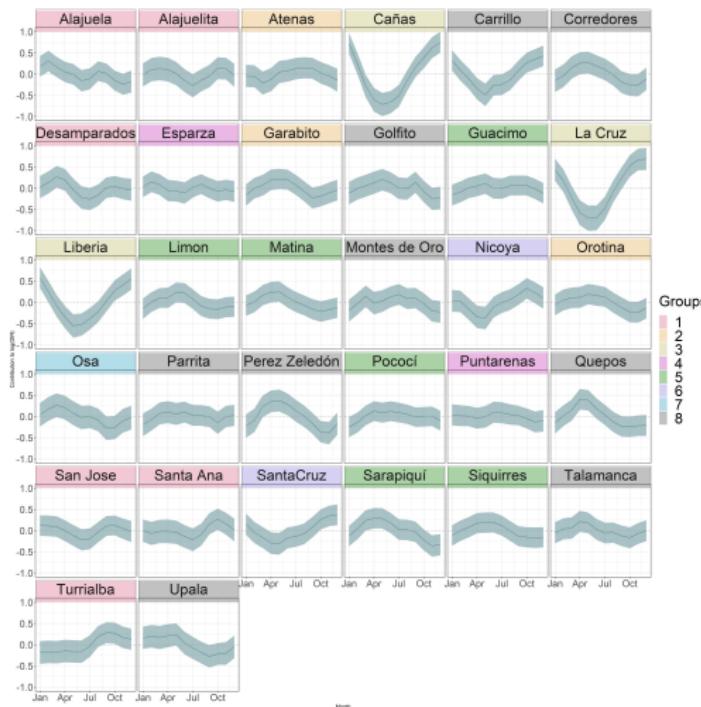
# Resultados

Cuadro 7: Métricas predictivas del conjunto de datos de entrenamiento y prueba del modelo seleccionado.

| Municipality       | Datos de entrenamiento |                   |            |                     | Datos de prueba |                     |            |                     |
|--------------------|------------------------|-------------------|------------|---------------------|-----------------|---------------------|------------|---------------------|
|                    | Independiente          |                   | CAR propia |                     | Independiente   |                     | CAR propia |                     |
|                    | NRMSE                  | NIS <sub>95</sub> | NRMSE      | NIS <sub>0,05</sub> | NRMSE           | NIS <sub>0,05</sub> | NRMSE      | NIS <sub>0,05</sub> |
| Parrita            | 1,24 × 10 <sup>9</sup> | 24483.9181        | 13.7622    | 9.8340              | 4.4446          | 46.3670             | 0.6773     | 7.9705              |
| Perez Zeledón      | 2.0103                 | 30.0358           | 0.8733     | 9.4268              | 1.2543          | 20.3380             | 0.1783     | 1.3942              |
| Pococí             | 2.2524                 | 12.7658           | 1.3001     | 6.1328              | 1.0524          | 12.0212             | 0.1284     | 1.8534              |
| Puntarenas         | 1.5802                 | 12.8416           | 0.9303     | 4.6933              | 1.3263          | 23.9665             | 0.1646     | 1.7646              |
| Quepos             | 56.8344                | 37.6499           | 10.0908    | 12.9868             | 36.9331         | 257.7509            | 1.1780     | 23.2153             |
| San Jose           | 0.2105                 | 12.6814           | 0.1328     | 5.2650              | 0.0215          | 1.2409              | 0.0155     | 2.3248              |
| Santa Ana          | 0.7837                 | 21.9311           | 0.5860     | 12.6213             | 0.6759          | 43.5264             | 0.3576     | 15.4854             |
| SantaCruz          | 35.1198                | 37.3329           | 8.4762     | 12.7070             | 6.5905          | 91.5896             | 0.3027     | 3.0635              |
| Sarapiquí          | 7.9218                 | 22.2392           | 2.8096     | 6.9807              | 0.3880          | 4.9820              | 0.1047     | 1.9942              |
| Siquirres          | 2.1184                 | 13.9371           | 1.4077     | 6.6140              | 2.7214          | 19.0646             | 0.3564     | 1.7977              |
| Talamanca          | 4.9193                 | 21.1751           | 2.2374     | 8.0522              | 0.1113          | 0.7642              | 0.7989     | 1.3152              |
| Turrialba          | 2.5905                 | 25.2386           | 1.9572     | 15.8268             | 1.0122          | 20.0694             | 0.2614     | 1.8489              |
| Upala <sup>1</sup> | 1.2744                 | 21.7159           | 0.9203     | 12.2963             | -               | -                   | -          | -                   |

<sup>1</sup> NRMSE y NIS<sub>0,05</sub> de los datos de prueba de Upala no se muestran porque los riesgos relativos observados son ceros.

# Resultados



**Figura 6:** Media posterior y intervalo creíble del 95 % de los efectos aleatorios mensuales de cada cantón.

# Resultados



Figura 7: Ilustración de 8 grupos con un comportamiento temporal similar.

# Resultados

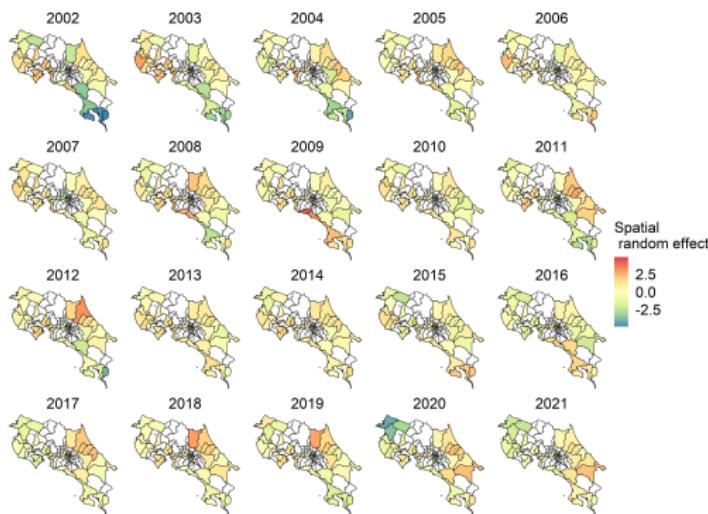
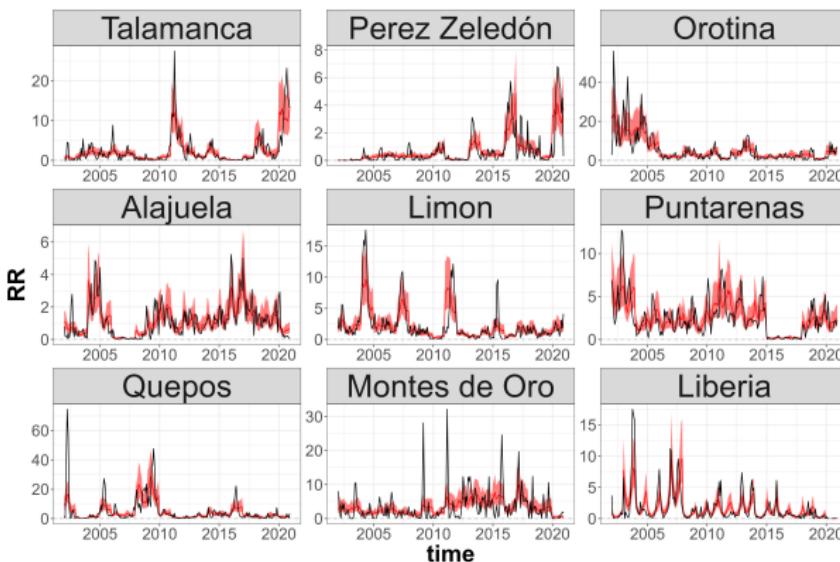


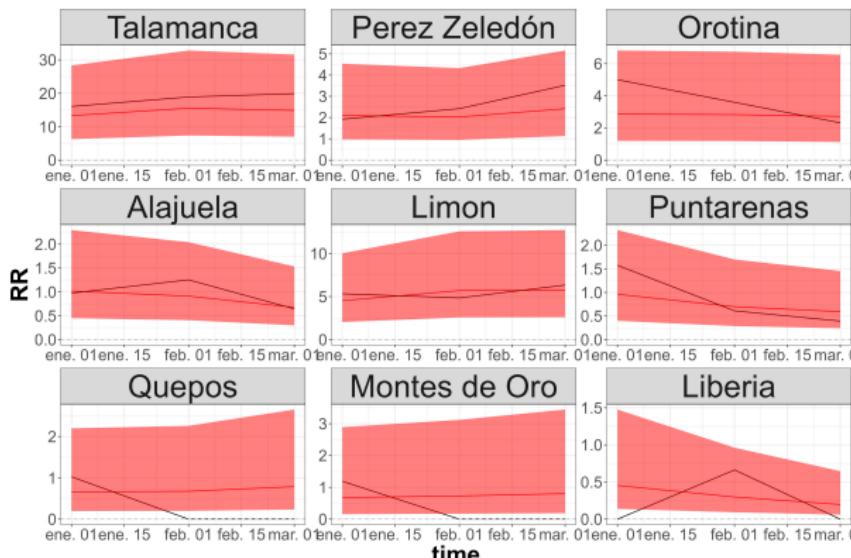
Figura 8: Contribución del efecto aleatorio espacial de cada año al log riesgo relativo del dengue.

# Resultados



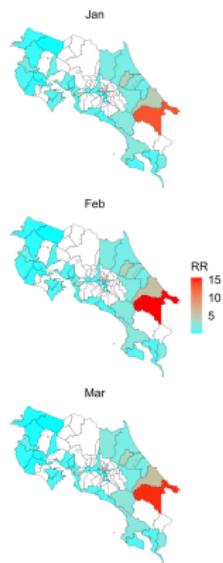
**Figura 9:** Comparación durante el período de ajuste. Los seis paneles superiores muestran los mejores cantones según la métrica NIS. Los tres paneles inferiores muestran los peores cantones según la métrica NIS. La línea roja representa el *RR* observado, mientras que la línea azul representa el *RR* estimado.

# Resultados

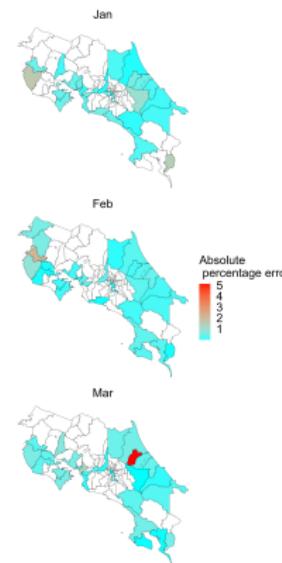


**Figura 10:** Comparación de pronósticos durante el período de prueba (2021). Los seis paneles superiores muestran los mejores cantones según la métrica NIS. Los tres paneles inferiores muestran los peores cantones según la métrica NIS.

# Resultados



(a) RR.



(b) Error porcentual absoluto.

**Figura 11:** Predicción del riesgo relativo y su error porcentual absoluto desde enero hasta marzo de 2021.

# Contenidos

## 1 Introducción

## 2 Modelos predictivos de riesgos de dengue con variables climáticas

- Entrenamiento del modelo
- Predicción
- Resultados
- Metodología
- Resultados

## 3 Emulador espacio-temporal climático (downscaling)

- Preliminares
- Metodología
  - Métodos de aproximación
- Resultados

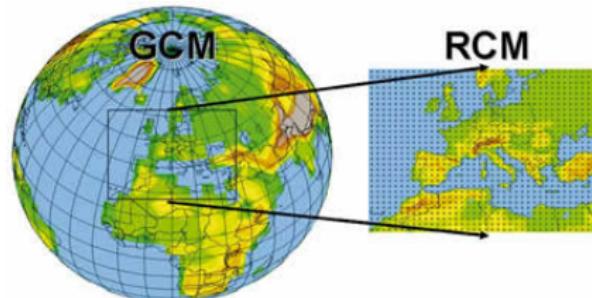
## 4 Conclusiones

# Preliminares

- Barboza, L. A., Chou-Chen, S. W., Córdoba, M. A. Alfaro, E. J., & Hidalgo, H. G (en prensa). **Spatio-temporal Downscaling Emulator for Regional Climate Models: a Comparative Study.** Environmetrics.
- Modelos climáticos como posible *input* para modelos epidemiológicos.

# Preliminares

- **Modelos Climáticos Regionales (RCM)**: describen la dinámica atmosférica y oceánica global. Modelos de reducción de escala que utilizan como entrada **un Modelo de Circulación General (GCM)**.
- Evaluación de los impactos del cambio climático y predicciones estacionales.
- Demanda computacional alta. (Wilby y Wigley, 1997)
- **Emulador estadístico**: aproximación de reducción de escala de la salida de RCM. (O'Hagan, 2006; Castruccio et al, 2014; Overstall y Woods, 2016)



# Objetivo

- Construir un emulador estadístico de reducción de escala de un modelo RCM, utilizando un enfoque espacio-temporal de coeficientes variables.
- Los resultados climáticos de mayor resolución son una barrera para muchas aplicaciones: modelos estadísticos/matemáticos epidemiológicos.

# Contenidos

## 1 Introducción

## 2 Modelos predictivos de riesgos de dengue con variables climáticas

- Entrenamiento del modelo
- Predicción
- Resultados
- Metodología
- Resultados

## 3 Emulador espacio-temporal climático (downscaling)

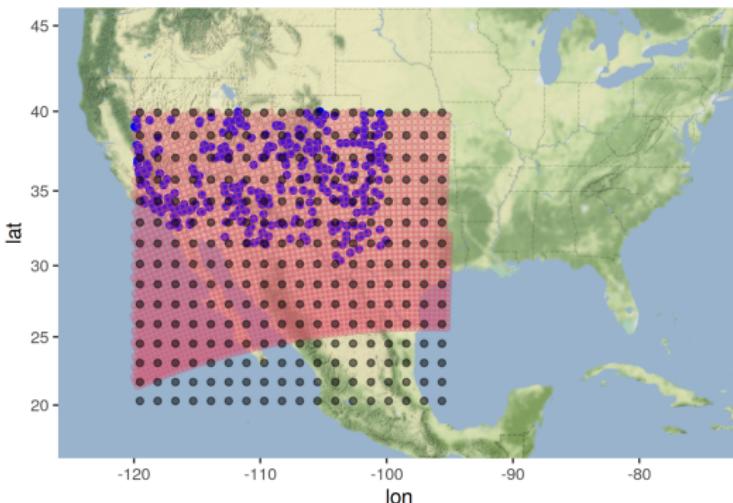
- Preliminares
- Metodología
  - Métodos de aproximación
- Resultados

## 4 Conclusiones

# Datos

- North American Regional Climate Change Assessment Program (NARCCAP).
- Generación de escenarios climáticos para su uso en investigación de impactos.
- **RCM:** Modelo Climático Regional Canadiense (CRCM):
  - Temperatura (grados Kelvin)
- **GCM:** Modelo de Sistema Climático Comunitario (CCSM):
  - Temperatura (grados Kelvin)
  - Velocidad vertical media de la presión estacional (PA/s) (OMEGA).
- **Datos observados:** Registros de temperatura del aire en superficie observados del the National Climatic Data Center (NCDC). Base de datos llamada como DSI-3200.
- Diferentes resoluciones espaciales. Resolución temporal común.

# Resolución de datos



- RCM (puntos rojos), GCM en la Región de América del Norte (puntos negros) y temperaturas observadas (puntos azules).
- Intersección entre el dominio de NARCCAP (Norteamérica) y el área de Monzón. Estudios futuros sobre el impacto del Monzón sobre América Central.
- 2482 puntos (RCM), 270 (GCM). Mensual en el tiempo.

# Comportamiento variando en el espacio

- Sea  $C_t(s), s \in \mathcal{S}$  la variable observada del Modelo Global y  $C_t(w), w \in \mathcal{W}$  la variable observada del Modelo Regional.
- Comportamiento de

$$Y_t(s) = \ln C_t(s) - \ln C_t(w)$$

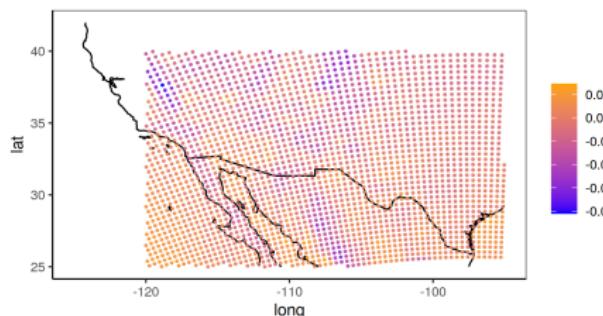
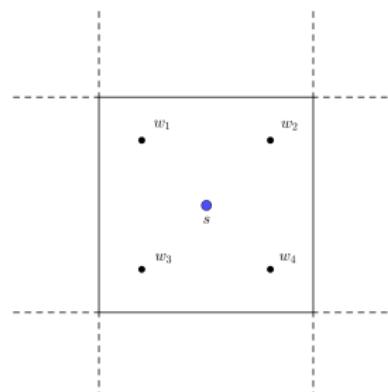


Figura 12: Diferencia de log-temperatura (regional vs global).

# Modelo emulador

- Dos conjuntos espaciales:  $\mathcal{S}$  (más grueso) y  $\mathcal{W}$  (más fino).
- Rejillas regulares en ambos conjuntos.
- Cada ubicación  $s \in \mathcal{S}$  es el centro de una región regular que contiene varios puntos sobre  $\mathcal{W}$ .



# Modelo emulador

- Modelo **Global**:

$$C_t(s) = \alpha + \beta' X_t(s) + \epsilon_t(s)$$

$\alpha$  y  $\beta$  son parámetros aleatorios y  $\epsilon_t(s)$  es ruido blanco en espacio y tiempo.

- Modelo **Regional**:

$$C_t(w) = [\alpha + \alpha_t^r(w)] + [\beta + \beta_t^r(w)]' X_t(s) + [\epsilon_t(s) + \gamma_t(w)]$$

donde

- $\alpha^r(\cdot) \sim N(\beta_0, \Sigma_0(\theta_0))$ ,
- $\beta^r(\cdot) \sim N(\beta_1, \Sigma_1(\theta_1))$  y
- $\gamma_t(\cdot) \stackrel{i.i.d}{\sim} N(0, \tau^2)$

y  $\Sigma_0$  y  $\Sigma_1$  son matrices de covariancia espacio-temporal separable.

# Modelo emulador

- Modelo espacio-temporal de coeficientes variables:

$$Y_t(w) := C_t(w) - C_t(s) = \alpha_t^r(w) + \beta_t^r(w)'X_t(s) + \gamma_t(w)$$

- Enfoque bayesiano:

$$\mathbf{Y}|\Phi \sim N(\beta_0 + \mathbf{X}^T\beta_1, \Sigma_Y)$$

- Evaluación del inverso y determinante de  $\Sigma_Y$  (que consume mucho tiempo). Consideraremos dos métodos de aproximación para evitar este problema.

# Métodos de aproximación

## ① Enfoque de Dambon (*varycoef*)

- Aproximación de la verosimilitud (Dambon et al., 2021).
- Independencia previa mutua en parámetros aleatorios.
- *tapering* en estructuras de covarianza.

## ② Integrated Nested Laplace Approximation (*INLA*)

- Aproximación Gaussiana de la verosimilitud bajo modelos más generales (modelos lineales generalizados con variables latentes). (Rue et al, 2009; Blangiardo et al, 2013.)
- Diseñado para reducir el tiempo de cálculo de modelos espaciales y/o temporales.
- Se hicieron varias simulaciones y se confirmó que INLA es más efectivo en tiempo de ejecución y precisión.

# Contenidos

## 1 Introducción

## 2 Modelos predictivos de riesgos de dengue con variables climáticas

- Entrenamiento del modelo
- Predicción
- Resultados
- Metodología
- Resultados

## 3 Emulador espacio-temporal climático (downscaling)

- Preliminares
- Metodología
  - Métodos de aproximación
- Resultados

## 4 Conclusiones

# Emulador para datos de NARCCAP

- $\mathcal{W}$ : resolución del RCM,  $\mathcal{S}$ : resolución del GCM.
- $C_t(\cdot) = \log T_t(\cdot)$ , donde  $T_t$  es la temperatura.
- $X_t(\cdot) = OMEGA_t(\cdot)$ . Tasa de ascenso vertical de las parcelas de aire.  
Proporciona una medida de los movimientos ascendentes y descendentes a gran escala en la atmósfera.
- Aproximación INLA con datos de entrenamiento de 1990-1998 (mensuales) y datos de prueba de 1999.
- Priors de Complejidad Penalizada (PC prior) para los parámetros autoregresivos y de varianza (efecto aleatorio).

# Emulador para datos de NARCCAP

**Modelo 0:** intercepto constante.

**Modelo 1:** Intercepto espaciotemporal variable que sigue un proceso aleatorio de ruido con la estructura de covarianza espacial sigue una Matern( $\nu = 1$ ).

$$\begin{cases} Y_t(w) &= \alpha_t(w) + \gamma_t(w) \\ \alpha_t(w) &= \epsilon_t(w) \end{cases}$$

**Modelo 2:** Intercepto espaciotemporal variable que sigue un proceso AR(1), donde la estructura de covarianza espacial sigue una Matern( $\nu = 1$ ).

$$\begin{cases} Y_t(w) &= \alpha_t(w) + \gamma_t(w) \\ \alpha_t(w) &= \rho\alpha_{t-1}(w) + \epsilon_t(w) \end{cases}$$

**Modelo 3:** Intercepto espaciotemporal variable que sigue un proceso AR(1) con la estructura de covarianza espacial sigue una Matern con una covariable:

$$\begin{cases} Y_t(w) &= \alpha_t(w) + \beta \cdot OMEGA_t(w) + \gamma_t(w) \\ \alpha_t(w) &= \rho\alpha_{t-1}(w) + \epsilon_t(w) \end{cases}$$

# Emulador para datos de NARCCAP

**Cuadro 8:** Comparación de modelos según las métricas predictivas (1990-1998: período de entrenamiento, 1999: período de prueba) y tiempo transcurrido en minutos.

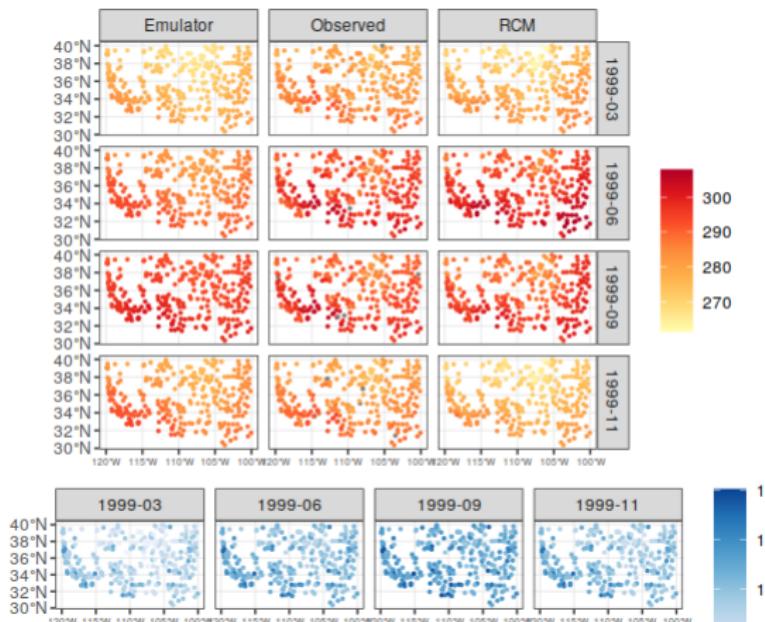
|              | RCM   | Modelo 0 |            | Modelo 1 |            | Modelo 2 |            | Modelo 3 |            |
|--------------|-------|----------|------------|----------|------------|----------|------------|----------|------------|
| Año          | MSE   | MSE      | $IS_{.95}$ | MSE      | $IS_{.95}$ | MSE      | $IS_{.95}$ | MSE      | $IS_{.95}$ |
| 1990         | 28.96 | 23.23    | 108.48     | 29.21    | 145.11     | 29.01    | 146.20     | 29.02    | 146.22     |
| 1991         | 32.07 | 18.98    | 96.64      | 32.16    | 153.97     | 31.94    | 155.17     | 31.93    | 155.18     |
| 1992         | 29.35 | 29.96    | 129.75     | 29.72    | 147.84     | 29.50    | 148.91     | 29.50    | 148.93     |
| 1993         | 28.28 | 28.87    | 128.90     | 28.77    | 138.62     | 28.55    | 139.78     | 28.56    | 139.82     |
| 1994         | 28.31 | 29.84    | 127.32     | 28.78    | 141.38     | 28.56    | 142.64     | 28.57    | 142.67     |
| 1995         | 23.56 | 24.49    | 110.77     | 23.86    | 128.43     | 23.67    | 129.60     | 23.66    | 129.61     |
| 1996         | 39.79 | 29.87    | 130.64     | 39.98    | 172.70     | 39.77    | 174.02     | 39.78    | 174.04     |
| 1997         | 23.88 | 20.76    | 97.22      | 24.29    | 128.93     | 24.10    | 129.99     | 24.10    | 129.99     |
| 1998         | 24.52 | 15.39    | 78.82      | 24.79    | 129.84     | 24.61    | 130.96     | 24.61    | 131.01     |
| 1999         | 29.41 | 38.38    | 149.91     | 37.85    | 36.52      | 40.32    | 40.42      | 40.78    | 41.17      |
| Tiempo ejec. | -     | 2.87     |            | 9.69     |            | 36.55    |            | 36.57    |            |

# Emulador para datos de NARCCAP

**Cuadro 9:** Estimaciones de parámetros e intervalo de predicción para el mejor escenario (Mod 1).

|           | L. inf. 95 % | Estimate   | L.sup. 95 % |
|-----------|--------------|------------|-------------|
| $\beta_0$ | -0.00238     | -0.00166   | -0.00094    |
| $\phi$    | 7.2622       | 7.4539     | 7.6463      |
| $\sigma$  | 0.0152       | 0.0155     | 0.0157      |
| $\tau$    | 0.00001977   | 0.00001989 | 0.00002001  |

# Emulador para datos de NARCCAP



**Figura 13:** Paneles superiores: temperaturas estimadas según el Modelo 1 (emulador), temperaturas observadas y RCM. Panel inferior: rango intercuartil para cuatro meses seleccionados durante el período de prueba. Los valores faltantes se muestran en gris.

# Contenidos

## 1 Introducción

## 2 Modelos predictivos de riesgos de dengue con variables climáticas

- Datos
- Modelo para cada cantón
  - Entrenamiento del modelo
  - Predicción
  - Resultados
- Modelo espacio-temporal
  - Metodología
  - Resultados

## 3 Emulador espacio-temporal climático (downscaling)

- Preliminares
- Metodología
  - Métodos de aproximación
- Resultados

## 4 Conclusiones

# Conclusiones

- Implementación de un modelo GAMLSS y RF utilizando diferentes variables climáticas para predecir el riesgo relativo de dengue en 32 cantones de mayor riesgo en este estudio.
- Modelos bayesianos espacio-temporales (INLA).
- La capacidad predictiva de estos métodos permite una vigilancia más exhaustiva y la identificación temprana de posibles brotes, lo que nos permite detectar brotes tempranamente, reduciendo el impacto social y económico.
- Trabajo futuro:
  - Más modelos espacio-temporales.
  - Más fuentes de datos como entrada (e.g. factores sociales)
  - Exploración de otros métodos de aprendizaje automático (NN y RNN).

# Conclusiones

- **Modelo jerárquico:**

$$\begin{cases} RR_t(w) & \sim C_t(w) + \text{Social/Climatic Factors}_t(w) \\ C_t(w) & = C_t(s) + \alpha_t(w) + \gamma_t(w) \\ \alpha_t(w) & = \rho\alpha_{t-1}(w) + \epsilon_t(w) \end{cases}$$

- Ventajas:

- Tiempo de computación: ajuste y muestreo.
- Propiedades de downscaling (reducción de escala).
- Pronóstico espacio-temporal.

# Muchas gracias por su atención!

Shu Wei Chou-Chen

email: [shuwei.chou@ucr.ac.cr](mailto:shuwei.chou@ucr.ac.cr)

website: <https://shuwei325.github.io/>



UNIVERSIDAD DE  
COSTA RICA

EEs  
Escuela de  
Estadística

CIMPA

Centro de Investigación en  
Matemática Pura y Aplicada

