MICHIGAN STATE UNIVERSITY

Mining Heterogeneous Time Series Datasets For Long Term Distribution Preserving Forecasts

Department of Computer Science & Engineering
College of Engineering

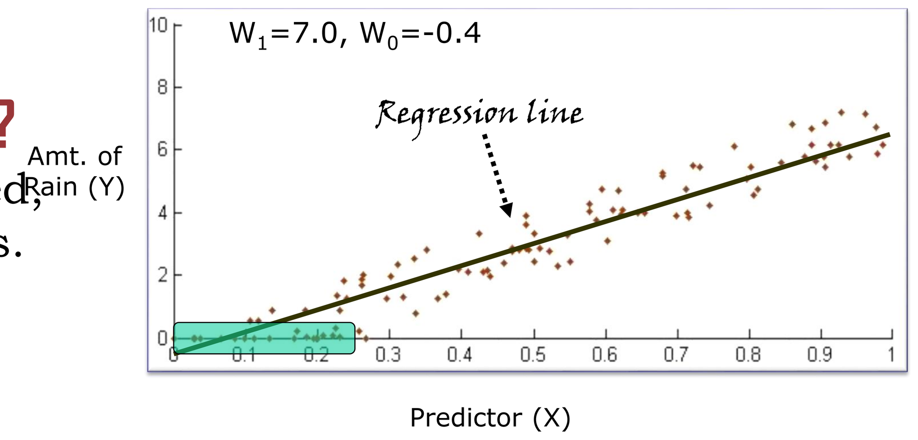
Zubin Abraham

Irregular Distributions

Applications such as climate modeling, frequently come across irregular distributions such as zero-inflated data that conventional forecasting approaches struggle to model.

What is Zero-Inflated Time series?

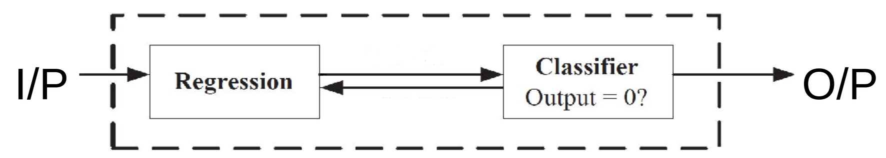
A time series that when discretized, has an abundant number of zeros. e.g.. Daily precipitation



What's the big deal?

Standard regression tends to underestimate the frequency of zeros and the magnitude of large non-zero values.

Proposed framework



ICR (Integrated Classification and Regression).

- Simultaneous classification and regression.

$$\arg\min_{\mathbf{w},\mathbf{y}} L(\mathbf{w},\mathbf{y}) = \sum_{i=1}^{n} c_i(c_i' - y_i y_i')^2 + T_1 \sum_{i=1}^{n}(y_i - c_i)^2 + T_2 \sum_{i,j=1}^{n} s_{i,j}[y_i y_i' - y_j y_j']^2 + T_3||w||^2$$

where,

$$y_i' = \sum_d w_d x_{i,d}, \quad y_i \in \{0,1\}$$

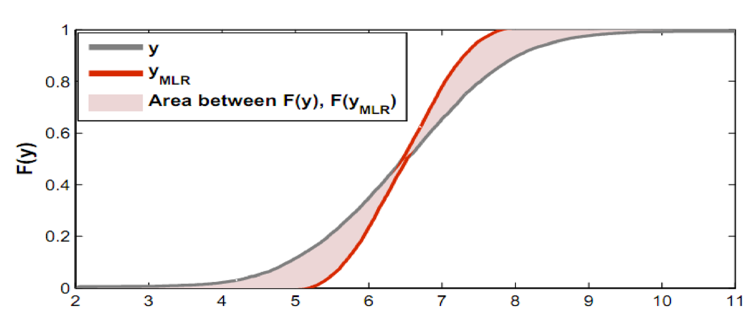
Shape of the Distribution

Conventional regression approaches minimize residual errors, but loose distribution shape.

Contour Regression (CR)

General framework for contour regression that combines

$$\min_{\beta} \sum_{i=1}^{n} (\gamma\pi(f(x_i), y_i) + (1-\gamma)\pi(f(x_i), y_{(i)}))$$



Multiple Linear Contour Regression (MLCR)

- Uses the ordinary least square (OLS) method.

$$\sum_{i=1}^{n} (\gamma(f(x_i,\beta) - y_i)^2 + (1-\gamma)(f(x_i,\beta) - z_i)^2)$$

Where,

$$f(X,\beta) = X\beta \text{ and } z_i = y_{(i)}$$

$$\Rightarrow \gamma(y - X\beta)^T(y - X\beta) + (1-\gamma)(z - X\beta)^T(z - X\beta)$$

$$\Rightarrow \hat{\beta} = (X^T X)^{-1}(\gamma X^T y + (1-\gamma)X^T z)$$

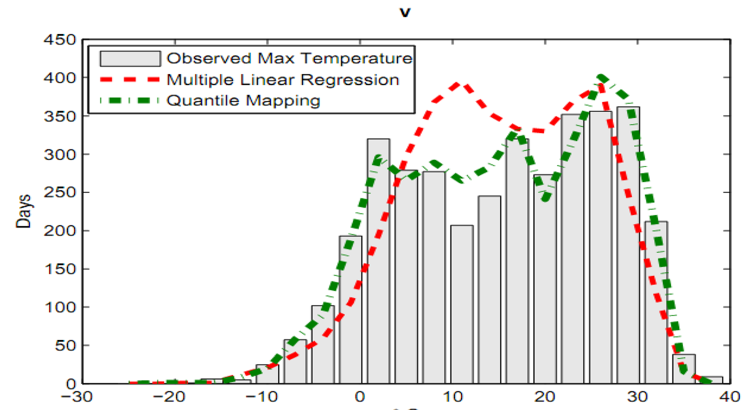
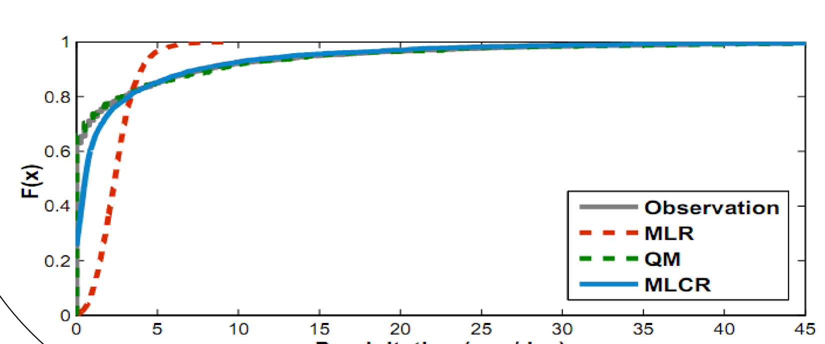


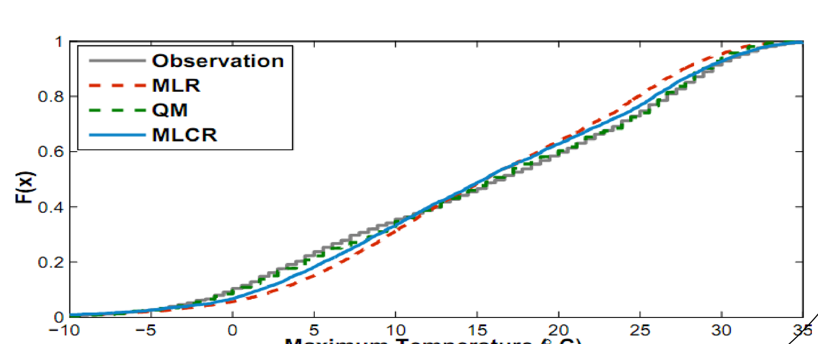
Table 2: Relative performance gain of MLCR over baseline approaches.

	RMSE % loss		RMSE-CDF % gain		RMSE-CDF win-loss %	
Dataset	MLR	Lasso	MLR	Lasso	MLR	Lasso
WRFG-T	1.9	1.7	39.0	41.7	100	100
CRCM-T	2.8	2.6	25.8	28.0	100	100
RCM3-T	2.0	1.8	35.3	39.2	100	100
WRFG-t	1.0	0.6	51.4	53.7	100	100
CRCM-t	1.9	1.6	38.2	40.1	100	100
RCM3-t	1.8	1.6	53.2	56.1	100	100
WRFG-P	28.8	28.3	74.3	75.8	100	100
CRCM-P	25.8	25.0	71.1	73.2	100	100
RCM3-P	29.9	29.5	75.6	76.7	100	100

Table 3: Percentage of stations that MLCR outperformed baseline in terms of σ and $\rho - CDF$

	σ win-loss%			$\rho - CDF$ win-loss%		
Dataset	MLR	Lasso	QM	MLR	Lasso	QM
WRFG-T	100	100	0	100	100	0
CRCM-T	100	100	0	100	100	0
RCM3-T	100	100	0	100	100	0
WRFG-t	100	100	0	78.6	85.8	64.3
CRCM-t	100	100	0	92.9	100	35.8
RCM3-t	100	100	0	92.9	85.8	85.7
WRFG-P	100	100	7.1	100	100	28.6
CRCM-P	100	100	0.0	100	100	50.0
RCM3-P	100	100	7.1	100	100	64.3





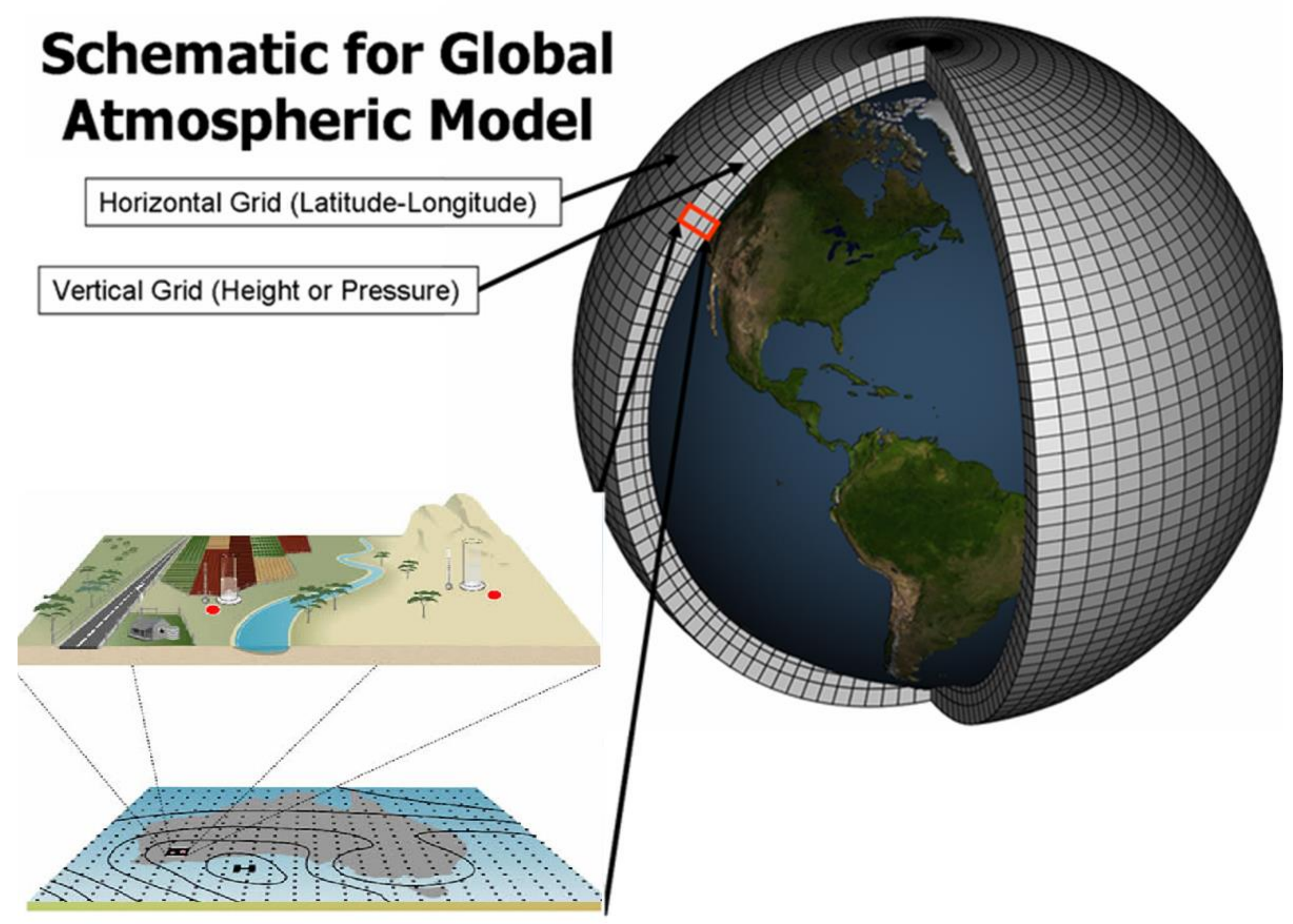
Overview

Long term forecasting is a multi-task problem, with emphasis on both accurate forecasting of individual data points as well as capturing the overall distribution characteristics of the response variable.

The unique selling point...

1. Handle irregular distributions.
2. Prioritize fidelity of extreme values forecasts.
3. Capture the shape (CDF) of the distribution, while minimizing residual errors.
4. Ensure associations and constraints are maintained in multi-output forecasting, while minimizing residual errors.

Schematic for Global Atmospheric Model



Acknowledgements

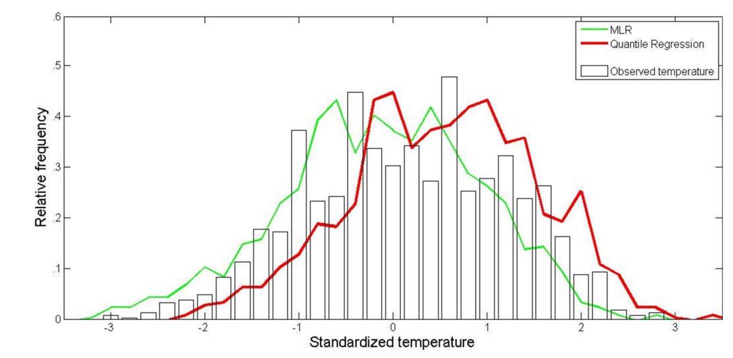
This work is partially supported by NSF grant III-0712987 and NASA award NNX09AL60G.

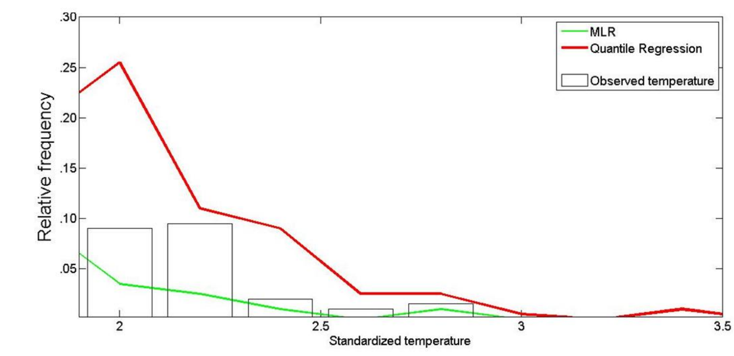
References

1. www.epa.gov/climatechange/science/futuretc.html
2. www.bom.gov.au/info/climate/change/gallery/65.shtml
3. **Abraham, Z** et al. *A Distribution Regularized Regression framework for Climate Modeling.* SDM'13
4. **Abraham, Z** et al. *Extreme Value Prediction for Zero-Inflated Data* –PAKDD'12,
5. **Abraham, Z** et al. *Smoothed Quantile Regression for Statistical Downscaling Of Extreme Events in Climate Modeling* - CIDU'11
6. **Abraham, Z** et al. *An Integrated Framework for Simultaneous Classification and Regression of Time-Series Data* –SDM'10
7. **Abraham, Z** et al. *A Semi-supervised Framework for Simultaneous Classification and Regression of Zero-Inflated Time Series Data with Application to Precipitation Prediction*- ICDM/SSTDM'09

Extreme values

Accurately forecasting extreme values is often very important. Unfortunately, conventional approaches that focus on accurately forecasting extreme values, fair poorly for the rest of the forecasted distribution.

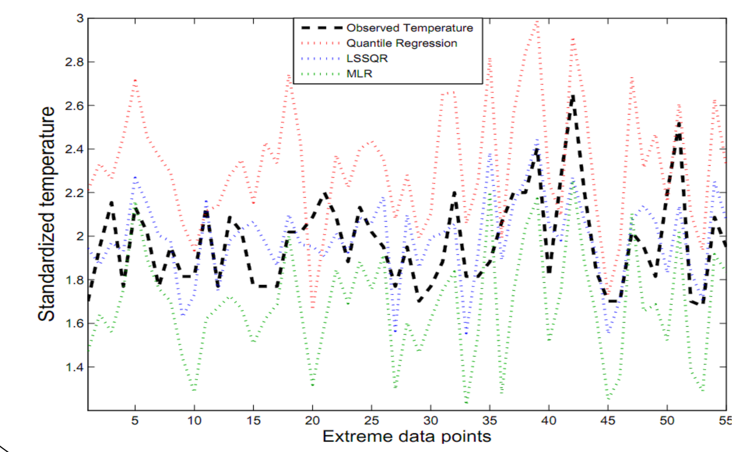


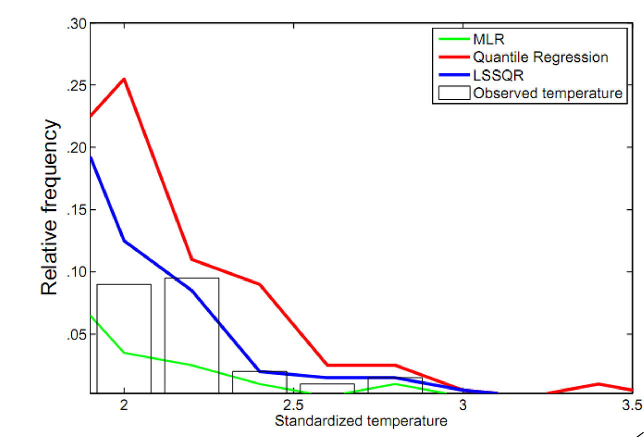


Linear Semi-Supervised Quantile Regression (LSSQR)

$$\arg\min_{\beta} \sum_{i=1}^{n} \rho_\tau(y_i - x_i^T\beta) + \lambda \sum_{i,j}^{n+m} w_{ij}(x_i^T\beta - x_j^T\beta)^2$$

$$\rho_\tau(u) = \begin{cases} \tau u & u > 0 \\ (\tau - 1)u & u \le 0 \end{cases}$$





Consistent Multiple-Output

Conventional single-output and multi-output regression approaches do not capture the association and constraints among multiple output variables.

Multi-Output Contour Regression (MCR)

$$\min_{\Omega} \sum_{i=1}^{N} (\gamma\Pi(h(\mathbf{x}_i,\Omega),\mathbf{y}_i) + (1-\gamma)\Pi(h(\mathbf{x}_i,\Omega),\mathbf{z}_{\hat{Y}Y}))$$

$$\hat{\mathbf{z}}_{\hat{Y}Y} = \mathbf{p}_{\hat{Y}}^{-k}(\mathbf{p}_Y^k(\mathbf{y}))$$

where,

$$\mathbf{p}_Y^k(\mathbf{z}) = \frac{1}{\kappa N} \sum_{i=1}^{N} \frac{\mathbf{p}_Y^{k-1}(\mathbf{z}) - \mathbf{p}_Y^{k-1}(\mathbf{y})}{\| \mathbf{p}_Y^{k-1}(\mathbf{z}) - \mathbf{p}_Y^{k-1}(\mathbf{y}) \|}$$

$$\mathbf{p}_Y^1(\mathbf{z}) = \frac{1}{\kappa N} \sum_{i=1}^{N} \frac{\mathbf{z} - \mathbf{y}'}{\| \mathbf{z} - \mathbf{y} \|}$$

$$MCR : \hat{\mathbf{z}}_{Y\hat{Y}} = \mathbf{p}_Y^{-k}(\mathbf{p}_{\hat{Y}}^k(h(\mathbf{x},\hat{\Omega})))$$

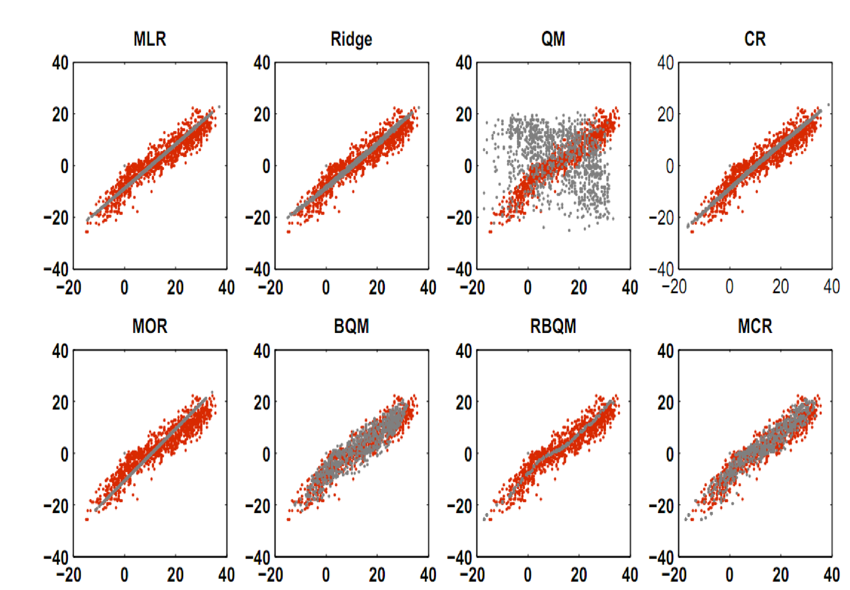


Fig. 5: Scatter plot portraying the fidelity of forecast values of various approaches replicating the observed associations among the bivariate temperature response variables.

Table 3: Performance of bivariate MCR over baseline approaches

Data set	RMSE: % of stations outperformed baseline			RMSE: Avg.improvement across stations over baseline			Kendall τ: % of stations outperformed baseline			Kendall τ: Avg.improvement across stations over baseline		
	MOR	QM	BQM	MOR	QM	BQM	MOR	QM	BQM	MOR	QM	BQM
$WRFG_1$	29	100	100	-0.06	0.18	0.17	64	100	100	0.03	0.40	0.41
$WRFG_2$	07	100	100	-0.08	0.16	0.16	79	100	100	0.04	0.38	0.39
$WRFG_3$	00	100	100	-0.07	0.31	0.30	0	100	100	-0.01	0.75	0.67
$CRCM_1$	93	100	100	0.06	0.25	0.25	100	100	100	0.13	0.52	0.53
$CRCM_2$	71	100	100	0.03	0.23	0.23	100	100	100	0.12	0.49	0.52
$CRCM_3$	07	100	100	-0.02	0.35	0.34	14	100	100	-0.01	0.78	0.73
$RCM3_1$	43	100	100	-0.02	0.20	0.20	79	100	100	0.06	0.46	0.46
$RCM3_2$	36	100	100	-0.03	0.19	0.18	79	100	100	0.06	0.47	0.45
$RCM3_3$	00	100	100	-0.07	0.31	0.30	0	100	100	-0.01	0.81	0.78

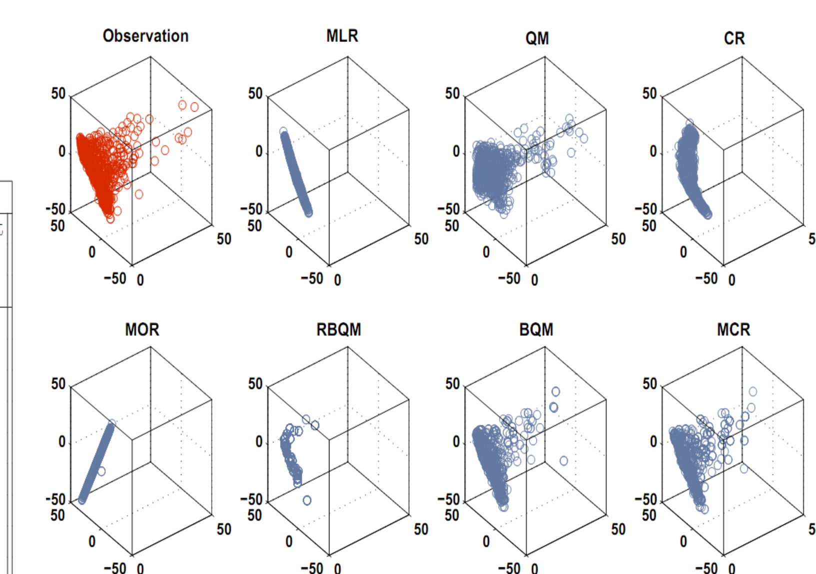


Fig. 6: Three dimensional scatter plot of the observed associations among maximum temperature, minimum temperature and precipitation as well as the respective forecasts made by the various single output and multiple output approaches.