Bias Correction

Asst.Prof.Dr.Jerasorn Santisirisombooon

Asst.Prof.Dr.Waranyu Wongseree

Ramkhamhaeng University

Definition of bias

The international definition of bias according to WMO is the correspondence between a mean forecast and mean observation averaged over a certain domain and time.

Causes

- Imperfect model representations of atmospheric physics
- Models are parameterized and evaluated on finite-length time series which may not cover the full range of atmospheric dynamics.
- The reference data sets (the “truth”) used for model parameterization and validation are inadequate.
- Incorrect initialization of the model or errors in the parameterization chain (GCMs).
- Incorrect boundaries provided by reanalyses or GCMs or inconsistencies between the physics of GCMs and RCMs (RCMs).

Bias Correction

advantages

- Provide realistic climate data
- Compare observed and simulated impacts during historical reference period
- Smooth transition into the future.

shortcomings

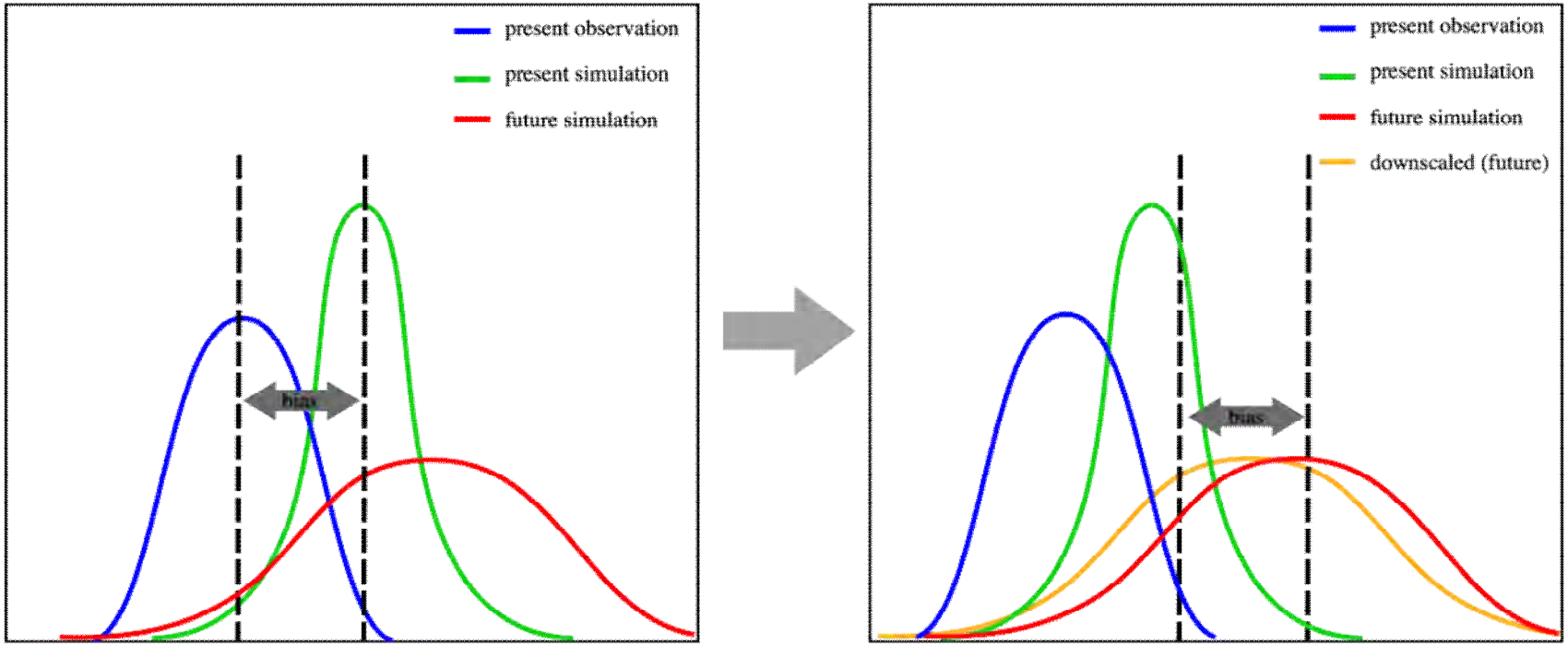
- Quality of the observations database limits the quality of the correction.
- It is assumed that the bias behaviour of the model does not change with time.
- Statistical bias correction often destroys the physical consistency of the different climate variables

Bias Correction

Methods to calibrate model simulations to ensure their statistical properties are similar to those of the corresponding observed values.

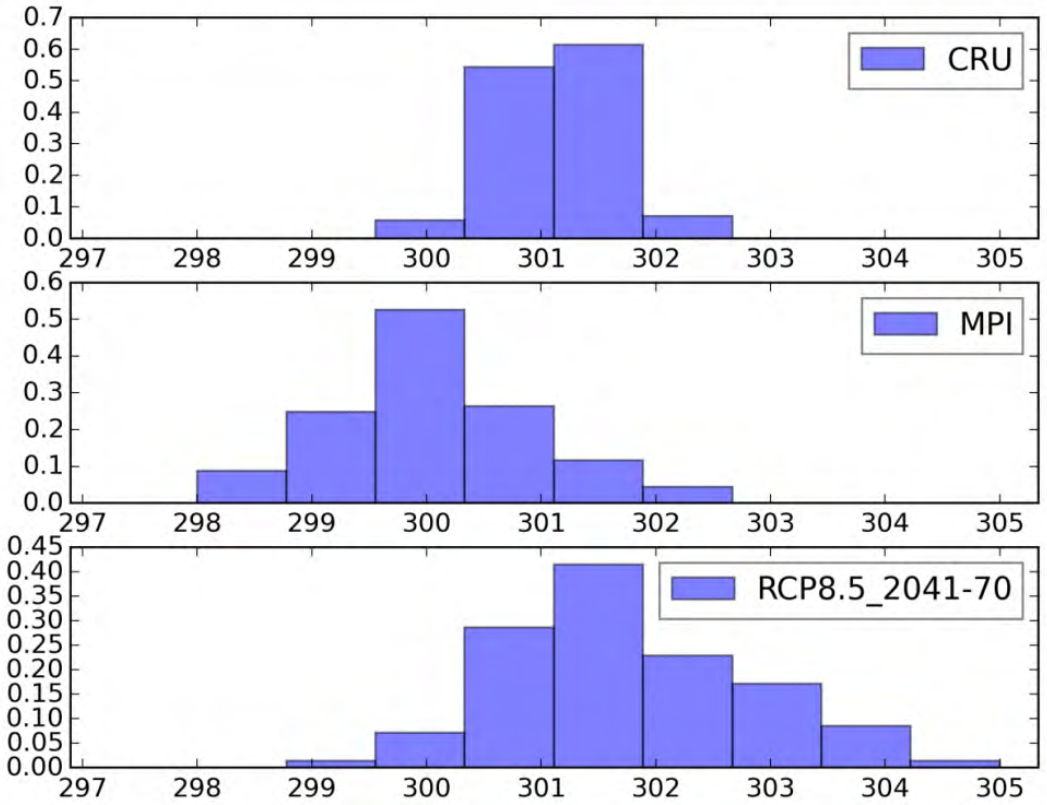
Bias correction is applied on each grid cell and for each time period separately to account for potential temporal and spatial structure in the biases.

Bias Correction Example



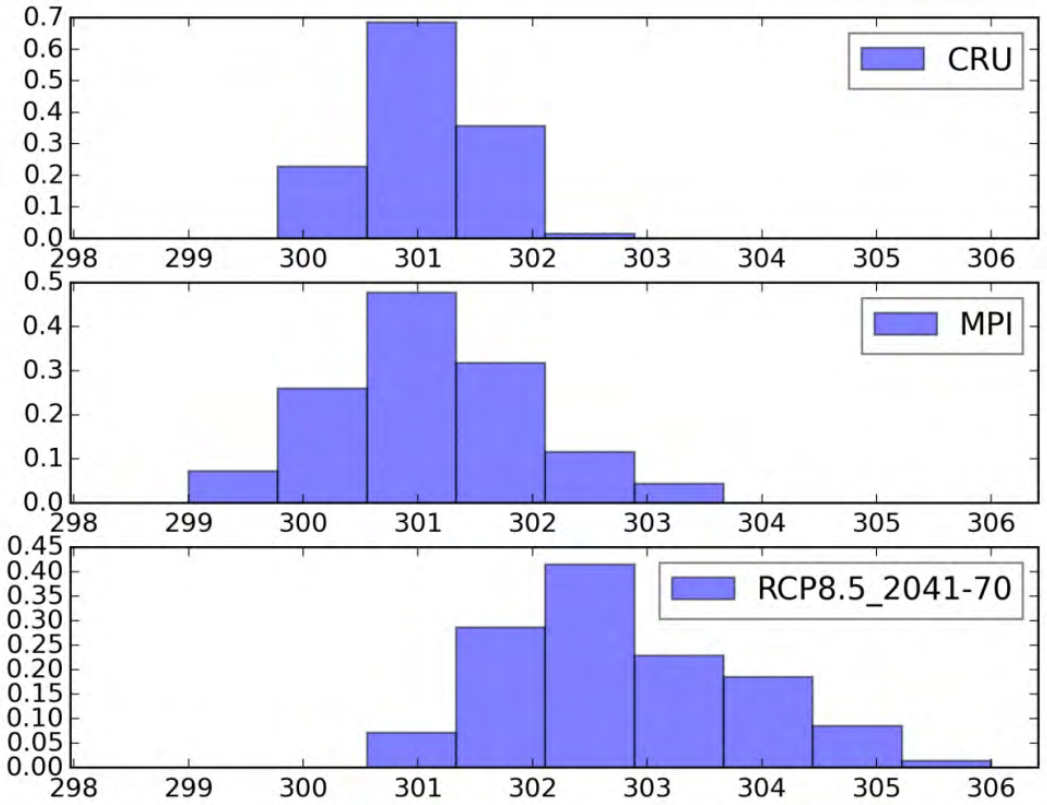
The mean bias of present simulation (green) from present observation (blue) is calculated and added to the future simulation [left diagram]. So the downscaled future simulation (yellow) has zero mean bias to the present observation [right diagram].

Original Distribution



Histogram of the original temperature distribution.

Corrected Distribution



The corrected data distribution using the Delta Correction Method.

Bias Correction Method

- Delta method (Additive method)
- Scaling method (Multiplicative method)
- Linear regression

Delta method

The formula for bias corrected data in projection period (x_{cor}) is

$$x_{cor} = a + bx_{pro}$$

where $a = \bar{x}_{obs} - \bar{x}_{sim}$ and $b = 1$

x_{pro} is the simulated data in projection period

$\bar{x}_{sim}$ is the means of simulated data in baseline period

$\bar{x}_{obs}$ is the means of observed data in baseline period

This method is preferable not to apply it to bounded variables (e.g. precipitation, wind speed, etc.) because values out of range could be obtained.

Scaling method

The formula for bias corrected data in projection period (x_{cor}) is

$$x_{cor} = a + bx_{pro}$$

where $a = 0$ and $b = \bar{x}_{obs} / \bar{x}_{sim}$

x_{pro} is the simulated data in projection period

$\bar{x}_{sim}$ is the means of simulated data in baseline period

$\bar{x}_{obs}$ is the means of observed data in baseline period

The scaling method is preferably applicable to variables with a lower bound (e.g. precipitation, because it also preserves the frequency).

Linear regression

The formula for bias corrected data in projection period (x_{cor}) is

$$x_{cor} = a + bx_{pro}$$

where a, b are obtained from least square of this equation

$$x_{obs} = a + bx_{sim}$$

x_{pro} is the simulated data in projection period

x_{sim} is the simulated data in baseline period

x_{obs} is the observed data in baseline period

Skill scores

A relative measure of the quality of the forecasting system compared to the benchmark or reference forecast

EXCEL Linear Regression Function

- EXCEL has a function that can be applied to analyze the (simple and multiple) linear relation of data base on least square method
- It is a kind of “ARRAY FORMULA”
- Syntax

  LINEST(known_y’s,known_x’s,const,stats)
- Output array : 5 x (m+1) where m is the number of independent variables.
- Press Control+Shift+Enter

	A	B	C	D	E	F	G
1	m_n	m_{n-1}	...	m_2	m_1	b	
2	se_n	se_{n-1}	...	se_2	se_1	se_b	
3	r^2	se_y					
4	F	df					
5	ss_{reg}	ss_{resid}					
6							
7	ss_{resid}	= $sum(y - y_{est})^2$		: the residual sum of squares			
8	ss_{total}	= $sum(y - y_{avg})^2$		: the total sum of squares			
9	ss_{reg}	= ss_{total} - ss_{resid}		: the regression sum of squares			
10							
11	r^2	= ss_{reg}/ss_{total}		: the coefficient of determination			
12							
13	The smaller ss_{resid} is, compared with the ss_{total}, the larger the value of r^2,						
14	which indicate of how well the equation resulting form the regression						
15	analysis explains the relationship among the variables						
16							
17	df	= n - k - 1		: degree of freedom			
18	n			: number of sample			
19	k			: number of independent variables			

	Name Box	B	C	D	E	F	G	H	I	J	K
21	F and df in LINEST output can be used to assess the likelihood of a higher F value occuring by chance. F										
22	can be compared with cirtical value in published F-distribution tables of EXCEL's FDIST can be used to										
23	calculated the probability of a larger F value occuring by chance.										
24											
25	Probability	= FDIST(F,v1,v2)		: probability that an F value this high occurred by chance							
26	v1	= n - df -1									
27	v2	= df									
28											
29	If FDIST(F,v1,v2) < alfa then we can conclude that F value this high is not occurred by chance										
30											
31	$F_{critical}$	= FINV(alfa,v1,v2)		: critical value of F							
32											
33	If F > $F_{critical}$ then we can conclued that F value this high is not occurred by chance.										
34											
35	Another hypothesis test will determine whetehr each slope coefficienti is useful in estimating the dependent variable (y).										
36											
37	se_n			: standard error of independent variable n							
38	t_n	$= m_n/se_n$		: t-observe value of independent variable n							
39											
40	If the absolute value of t_n is sufficient high, it can be concluded that the slope coefficient mn is usefuel in estimating y.										
41											
42	$t_{critical}$	=TINV(alfa,df)		: t critical, two tailed.							
43											
44	If the absolute value of t_n is greater than $t_{critical}$, n is an important variable when estimating y.										