Environmental Intelligence using Machine Learning Applications

N. Malakar

Journal of Nepal Physical Society
Volume 8, No 2, 2022
(Special Issue: ANPA Conference 2022)
ISSN: 2392-473X (Print), 2738-9537 (Online)

JNPS, **8** (2), 42-47 (2022)
DOI: http://doi.org/10.3126/jnphyssoc.v8i2.50148

Published by: Nepal Physical Society
P.O. Box: 2934
Tri-Chandra Campus
Kathmandu, Nepal
Email: nps.editor@gmail.com

ANPA Conference 2022
ISSN: 2392-473X(Print), 2738-9537 (Online)

Review Article ©Nepal Physical Society
DOI:http://doi.org/10.3126/jnphyssoc.v8i2.50148

Environmental Intelligence using Machine Learning Applications

N. Malakar[a)]

Worcester State University, Worcester, MA 01602, USA

[a)]*Electronic mail: nmalakar@worcester.edu*

Abstract. The big data deluge has presented us with a unique opportunity to observe the environment. These sensors employ the basic physics principles in sensing the environment. These include but are not limited to, remote sensing of air quality, temperature, or other biophysical variables. Although a great effort has been placed into collecting the data, a greater effort must be placed into their societal applications. Machine Learning tools can provide easy access to build such applications. Continuous monitoring and alerting the interested parties can prevent some undesirable outcomes. For example, a weather forecast is widely used to predict the temperature/precipitation a few days in advance. Similarly, new applications can be developed to practice intelligent decision-making that affects public health. Recent progress in air quality studies is promising to develop such environmental intelligence. In this review article, we illustrate the use of Machine Learning in making predictions and discuss some of the applications of the relevant data sets for Environmental Intelligence.

Received: 31 August 2022; **Accepted**: 04 December 2022; **Published**: 31 December 2022

Keywords: Machine Learning; Remote Sensing; Environmental Physics

INTRODUCTION

At the heart of environmental sensing lies the basic principles of physics. This could be done using in-situ observation of air pollution, temperature, other physical and meteorological variables, etc. Instrumentation used to be expensive – up to the extent that some of the in-situ measurements were cost prohibitive. However, recent advances in low-cost sensor technology, embedded devices connected to the internet, and development in modern remote sensing instruments have enabled a deluge in data produced through these sensors. This presents us with a unique opportunity to gather information about the environment. However, the enormous volume, variety, velocity, and veracity of the data can present challenges in getting useful insight. This paper briefly reviews the opportunities and challenges in the intelligence of environmental data-set using various tools such as machine learning applications. We will briefly discuss the importance of Earth observation and the basic idea of machine learning and Environment Intelligence with potential applications.

EARTH OBSERVATION

The scientific method involves the process of hypothesis formulation, experimental design, data collection, and analysis of the parameters defining the hypothesis. This can ultimately lead to a more refined set of hypotheses or a new set of model parameter values. One of the exciting pieces of information that can be extracted from such a process is an insight into the interconnection between the relevant variables in a system. With the same view, we can consider our planet Earth to be a system of significant interconnected variables. In an attempt to understand the variables' interconnection, we can imagine a vast model that can be constructed to replicate the general circulation and regulation of other physical variables that can be observed. However, the best model of the Earth is the Earth itself. The universe is a laboratory where the observation data can provide clues to the interactions of the variables.

A simple model illustrating energy flow is represented as shown in Fig. 1 (a), where the energy is transferred from high to low potential through a resistor (R). Suppose we represent the problem as a circuit, with the potential difference represented as a battery and the resistor as a light-emitting diode (LED) (Fig. 1(b)). In that

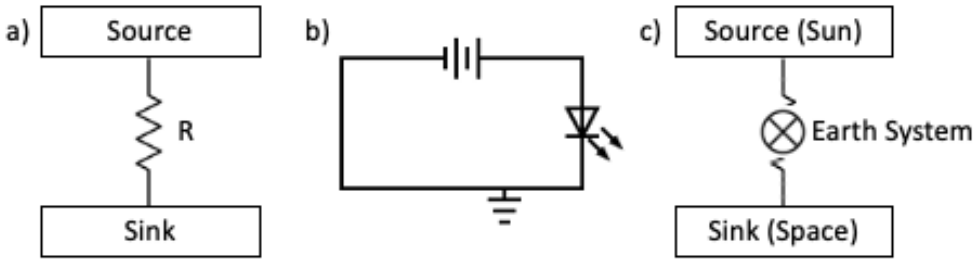

FIGURE 1. (a) As an example of a simple model, a source is connected to the sink through a resistor, which could be illustrated by using a circuit diagram (b) where the interaction gives rise to the illumination of the LED. (c) The radiation energy balance on Earth can illustrate a zeroth-order energy balance model for Earth.

case, the relationship between the resistance (R), current (I), and potential difference (V) can be represented using Ohm's law $V = IR$. Similarly, a zeroth order energy balance model for Earth can be illustrated by the radiation energy balance on the earth in Fig. 1(c). A simple calculation using Stefan-Boltzmann law and the zeroth order energy balance model shows that the surface temperature of the Earth would be

$$T_s = \left(\frac{S(1-\alpha)}{\sigma}\right)^{0.25} \tag{1}$$

where S denotes the solar constant, α represents the average albedo of the earth's surface, and σ is the Stefan-Boltzmann constant. A simple physical calculation shows the earth's temperature would have been -18°C. However, our observations do not agree with it. This requires updating our simple model based on physical reality- the Greenhouse gases provide a warm blanket to make the planet habitable. The updated model incorporates the difference between incoming and outgoing radiative fluxes $R\downarrow - R\uparrow$, equated to the heat energy $mc\Delta T$:

$$(R\downarrow - R\uparrow)\, a\Delta t = mc\Delta T, \tag{2}$$

where $R\downarrow$ and $R\uparrow$ represent the incoming and outgoing energy terms. The first two terms on the left-hand side have the unit of energy density, W/m^2, which has been multiplied by the surface area a of the earth and time (Δt) to convert it into the dimensions of energy; ΔT represents the temperature, m the mass and c is the specific heat capacity of the system. This can be rearranged to write the temperature evolution equation as:

$$\{0.25(1-\alpha)S - \varepsilon\sigma T^4\} = c\frac{\Delta T}{\Delta t}, \tag{3}$$

where $0 < \varepsilon < 1$ represents the emissivity of the surface. With $S = 1170W/m^2$, $\alpha = 0.3$, $\varepsilon = 0.6$ and $\sigma = 5.67\times10^{-8}W/m^2/K^4$, the surface temperature is approximately 14°C. This value comes to close agreement with the global average temperature. This model can further be developed to incorporate many other complex parameters. Next, we focus on atmospheric components and their observation as they will play a critical role in the Earth's energy balance.

Earth system observations can be made using in-situ or remote sensing observation. Remote sensing data from Earth observation provide a continuous set of data collected from space. Establishing and maintaining ground-station data at every point on earth is tedious and cost-prohibitive. The resolution of the desired model can dictate such placements. If we want high-resolution sensor networks, it can be cost-prohibitive. On the other hand, remote sensing sensors are spaceborne sensors pointed toward the earth that can cover a larger area providing a continuous stream of data. Analysis of such a volume of data can be conveniently considered Big Data. For example, Moderate Resolution Spectroradiometer (MODIS) instruments onboard the NASA Terra and Aqua satellite platforms measure various parameters such as aerosols, ocean color, atmospheric water vapor, cloud properties, surface and temperature, different atmospheric parameters, etc. These parameters are observed as remote sensing images and encoded as profiles on about 250 m to 1 km pixel resolutions. These suites of sensors have been traversing the surface continuously since their launch in 1999 (Terra) and 2002 (Aqua), respectively. Currently, more than two dozen satellites operated by NASA constantly observe the Earth's parameters.

These Earth observation data mount upwards of 100s petabytes of data in the archival system. Therefore, we can safely say that these Earth observation data serve as big data. A proper analysis presents both the challenge and opportunity for curious minds. The deluge comes with challenges and opportunities at the same time. While the appropriate analysis and inference of the data can be regarded as a daunting task, it also provides a chance to understand the interrelation between the variables.

MACHINE LEARNING

Machine Learning (ML) is an emerging field that has proven its usefulness in a wide variety of applications in science, healthcare, business insights, and engineering. ML involves data-based learning. These applications are helpful, especially when the model may not be fully parameterized. I do agree with the sentiment that a first-principal theory and approach, as illustrated above, for the energy balance model would be much appreciated. However, in the complex world of interconnected variables, some of these relations could be non-linear, and some of the variables could be hidden as well. Therefore, non-parametric, non-linear methods such as ML deserves careful consideration.

The ML technique follows the basic scientific process: formulating hypotheses, testing, and refining hypotheses. Developing a hypothesis with a defined set of variables can be tricky, especially if you miss the relevant variables in the formulation. For example, applying the equations of type $y = mx + b$, where the variables y, m, x and b represent the Force, mass, acceleration and a constant additive, respectively. This linear equation may work well for an ideal case, but may not work well for a problem if you needed to include the extra factors such as the effect of friction. However, suppose there could be an objective set of tools for automating the discovery of such variables. In that case, one could possibly model the observation data sets without a defined set of fixed parameters.

Depending on the process involved, the ML techniques can be divided into three primary types : Supervised, Unsupervised, and Reinforcement learning methods. The Supervised ML techniques are routinely used for regression and classification problems. The regression methods are used in estimating and predicting the relationship between variables, especially when the dependent variables are continuous, while classification is used to build a model with discrete or categorical variables. Examples of regression may include interpolation and extrapolation of data points, while classification of leaf-types are examples discrete supervised techniques based on the sample data. On the other hand, Unsupervised ML techniques are used when the output is not known for the available input samples. Unsupervised ML techniques provide an insightful tools to explore the patterns in data. Examples could include clustering of images, based on their statistical features. Meanwhile, the Reinforcement learning is a feedback-based ML techniques, where the algorithm is assigned with a cost function and is rewarded for iteratively learning from the environment. Examples could include These ML techniques are data-based: essentially, ML algorithms work by searching through a large space of candidate hypotheses to optimize the learning outcomes defined by the performance metric.

The ML process is also regarded as a data-intensive method. In the past, big data presented an obstacle in processing the training datasets. However, recent progress in low-cost computational machines and cheaper storage has driven rapid progress in ML applications because data processing has become relatively more affordable. Next, we will present a simple connected problem that may resemble some physical systems.

A Circuit Example

We present a simple circuit model as an example of how we could use ML programs to explore the relationship between variables. This example is intended to be presented as a toy model. So, the readers are advised to consider it carefully.

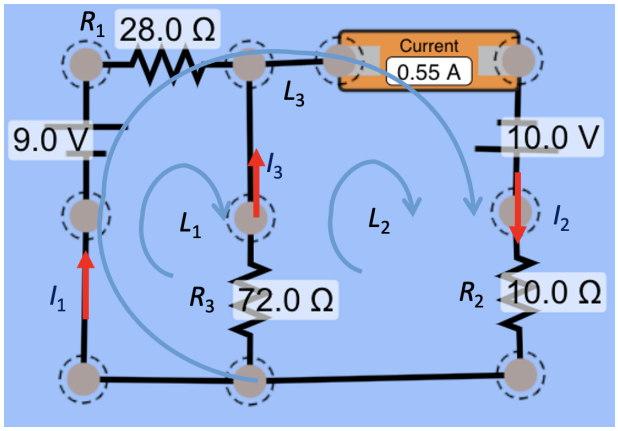

FIGURE 2. An example circuit. In this example, we simulate the circuit to estimate the currents using training data only using Machine Learning. The circuit is constructed using the PhetSim software for demonstration.

Let us consider a mesh network as shown in Fig. 2, where arrows indicate the currents on the circuit. If one were to solve for the circuit, Kirchhoff's law could be applied to solve for the three currents (I_1, I_2, I_3) given the voltages and resistance. This is possible using Ohm's law and concepts in the conservation of charge and energy. The circuit current values can be found by using a forward problem:

$$\begin{pmatrix} I_1 \\ I_2 \\ I_3 \end{pmatrix} = \frac{1}{(R_1R_2 + R_2R_3 + R_1R_3)} \begin{pmatrix} R_3V_1 + & R_3V_2 + & R_2V_1 \\ R_3V_1 + & R_1V_2 + & R_3V_2 \\ R_1V_2 - & R_2V_1 & \end{pmatrix} \tag{4}$$

However, in this example, we want to solve the same problem without using Equation (4). A challenge is to predict the observation of circuit current values (I_1, I_2, I_3) given the circuit potential (V_1, V_2), and resistor combinations (R_1, R_2, R_3) while assuming nothing else. Kirch-

hoff's laws are an example of establishing a model without knowing physics.

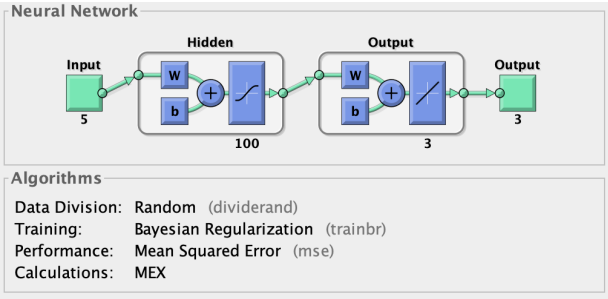

FIGURE 3. Neural Network set up with five input variables (Resistors and Potential difference in Fig. 2 used to predict the three currents in the circuit.

To prepare the training dataset, an array of 4000x5 random values $\{V_1, V_2, R_1, R_2, R_3\}$ were selected as input to simulate the output currents $\{I_1, I_2, I_3\}$ using Eq. 4. The resistor values were selected from the range of $0 : 200\Omega$, while the potential differences were set to range from 0 to 5 volts. An artificial neural network was created with five input variables, 100 hidden layers, and three outputs (Fig 3, and trained for 1000 epochs. The data were divided into random sets of training (60%), testing (30%), and validation (10%) during the training using Bayesian Regularization, and the performance of the algorithm was monitored using the Mean Squared Error metric. When the training was completed, we then used the neural network model to predict a thousand sets of unseen voltage and resistor combinations $\{V_1, V_2, R_1, R_2, R_3\}$ to predict the output currents. Since this was a simulation experiment, we could calculate the estimated currents to verify that our results agreed with the simulated outputs. This serves as the benchmark test dataset to check the efficacy of our model.

To test the efficacy of the model, we tested 1000 random circuits with the potential and resistor combinations, and predicted the currents. Figure 4 shows the residue (Predicted - True values) on these 1000 test observation datasets. The low difference between the predicted and observed data shows that the model performance is accurate within the milli-Amperes range. An example of Kirchhoff's law has now been presented without telling the machine the fundamental physics of current, voltage, and resistors. Although it is a simple example, the possible power of data-based methods in exploring the interconnected world is illustrated.

CONNECTING DOTS

In the previous two sections, an overview of Machine Learning and climate feedback system was introduced.

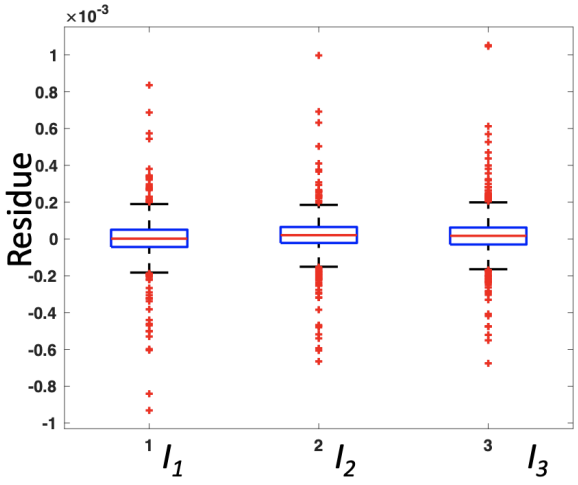

FIGURE 4. The residue between known current values and estimated current values shows that the difference is in the milli Amperes range.

Next, I would like to draw a connection between these two independent topics.

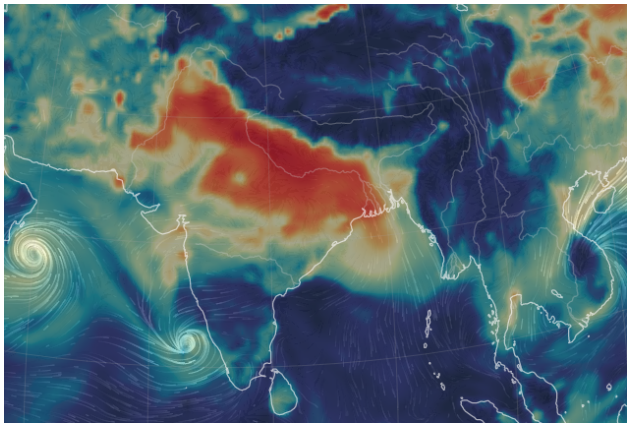

FIGURE 5. Map of global surface wind shows stalled air pollution due to air vortex in the Indian Ocean during November 2019. The extent of air pollution is from blue (low) to red (higher pollution). Data courtesy of @EarthWindMap.

ENVIRONMENTAL INTELLIGENCE

Physicists are known for training in extracting hidden patterns based on observation and data sets. The movement for gradual model development, testing, and validation help extract insights from big data sets. As discussed in the previous sections, some relationships could be complex, non-linear, and guided by "hidden" processes. Environmental data sets are derived from a live system where

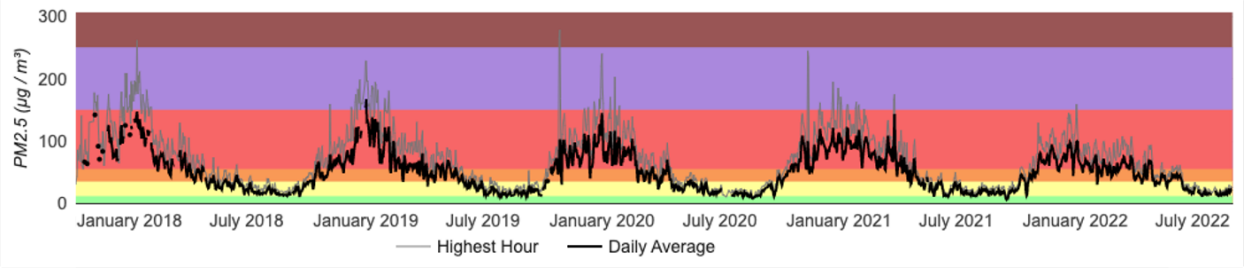

FIGURE 6. PM 2.5 daily dataset. Data courtesy of BerkeleyEarth.

all the variables tightly interact. Getting insights from these datasets is not a simple task. The first step towards intelligence is to be able to ask the most relevant questions and seek out the relevant variables. Then, one can look for patterns or trends and be able to predict the variable's behavior.

Air pollution can be an example challenge that could benefit from Environmental Intelligence. Fig. 5 shows the persistent wind pattern that dictated the distribution of pollutants. The question to ponder is whether we can we set up an intelligence system with the model and observation to predict its occurrence.

Figure 6 shows Indian subcontinent's average daily PM2.5 data set from January 2018 through July 2022. Although, a noisy dataset, due to daily variation in the air quality, it is easy to extract the information about the yearly cycles with a distinct repetition. As shown in the figure, the color codes represent the PM2.5 air pollution concentration as it implies that sensitive groups could have an adverse impact on their health. The average air pollution during the winter seasons reaches 100 $\mu g/cm^3$– a harmful level with severe health consequences. However, if we want an automated detection of the yearly patterns that human eyes can easily track from the data sets like these need a machine learning system. Since the air pollution level exceeds the safe level for human exposure, automated discovery of such patterns may help develop policies for minimizing human exposure. More importantly, the impacts on climate variability due to air pollution present another unique opportunity for the issue of Environmental Intelligence.

CONCLUSIONS AND DISCUSSIONS

Recent advances in Artificial Intelligence and Machine Learning technologies have given us a unique opportunity to combine the skill-sets of physics, statistical data analysis, and remote sensing with improving life quality. The big data deluge invites us to explore our world with a new vision. The ML tools can provide new ways of exploring the inter-relationship between the variables that may not be linked yet using the first principles. This can benefit many areas of science. Especially informed decision-making and data-based policy development seem poised to take advantage of these advances.

EDITORS' NOTE

This manuscript was submitted to the Association of Nepali Physicists in America (ANPA) Conference 2022 for publication in the special issue of the Journal of Nepal Physical Society.

REFERENCES

1. C. M. Bishop *et al.*, *Neural networks for pattern recognition* (Oxford university press, 1995).
2. L. Breiman, J. Friedman, R. Olshen, and C. Stone, "Classification and regression trees (the wadsworth statistics/probability series) chapman and hall," New York, NY , 1–358 (1984).
3. C. Cortes and V. Vapnik, "Support-vector networks," Machine learning **20**, 273–297 (1995).
4. H. B. Demuth, M. H. Beale, O. De Jess, and M. T. Hagan, *Neural network design* (Martin Hagan, 2014).
5. J. Friedman, T. Hastie, and R. Tibshirani, "The elements of statistical learning, volume 1 springer series in statistics springer," (2001).
6. D. J. Lary, G. K. Zewdie, X. Liu, D. Wu, E. Levetin, R. J. Allee, N. Malakar, A. Walker, H. Mussa, A. Mannino, *et al.*, "Machine learning applications for earth observation," Earth observation open science and innovation **165** (2018).
7. S. B. Kotsiantis, I. Zaharakis, P. Pintelas, *et al.*, "Supervised machine learning: A review of classification techniques," Emerging artificial intelligence applications in computer engineering **160**, 3–24 (2007).
8. D. J. Lary, F. S. Faruque, N. Malakar, A. Moore, B. Roscoe, Z. L. Adams, and Y. Eggelston, "Estimating the global abundance of ground level presence of particulate matter (pm2. 5)," Geospatial health **8**, S611–S630 (2014).
9. S. Osowski and K. Garanty, "Forecasting of the daily meteorological pollution using wavelets and support vector machine," Engineering Applications of Artificial Intelligence **20**, 745–755 (2007).
10. V. N. Vapnik, "The nature of statistical learning," Theory (1995).
11. V. Vapnik, *Estimation of dependences based on empirical data* (Springer Science & Business Media, 2006).
12. T. M. Smith and V. Lakshmanan, "Real-time, rapidly updating severe weather products for virtual globes," Computers & Geosciences **37**, 3–12 (2011).

13. D. J. Lary, L. O. H. Wijerante, G. K. Zewdie, D. Kiv, D. Wu, F. S. Faruque, S. Talebi, X. Yu, Y. Zhang, E. Levetin, *et al.*, "Machine learning, big data, and spatial tools: A combination to reveal complex facts that impact environmental health," in *Geospatial Technology for Human Well-Being and Health* (Springer, 2022) pp. 219–241.
14. N. K. Malakar, D. J. Lary, and B. Gross, "Case studies of applying machine learning to physical observation," in *AGU Fall Meeting Abstracts*, Vol. 2018 (2018) pp. H31H–1978.
15. N. K. Malakar, D. J. Lary, A. Moore, D. Gencaga, B. Roscoe, A. Albayrak, and J. Wei, "Estimation and bias correction of aerosol abundance using data-driven machine learning and remote sensing," in *2012 Conference on Intelligent Data Understanding* (IEEE, 2012) pp. 24–30.
16. C. Arizmendi, J. Sanchez, N. Ramos, and G. Ramos, "Time series predictions with neural nets: application to airborne pollen forecasting," International journal of biometeorology **37**, 139–144 (1993).
17. J. D. Berman, P. N. Breysse, R. H. White, and F. C. Curriero, "Health-related benefits of attaining the daily and annual pm2. 5 air quality standards and stricter alternative standards," in *A103. HEALTH SERVICES RESEARCH AND ADMINISTRATIVE DATABASES* (American Thoracic Society, 2012) pp. A2317–A2317.
18. M. Castellano-Méndez, M. Aira, I. Iglesias, V. Jato, and W. González-Manteiga, "Artificial neural networks as a useful tool to predict the risk level of betula pollen in the air," International Journal of Biometeorology **49**, 310–316 (2005).
19. T. Mitchell, B. Buchanan, G. DeJong, T. Dietterich, P. Rosenbloom, and A. Waibel, "Machine learning," Annual review of computer science **4**, 417–433 (1990).
20. N. Malakar, D. Lary, D. Gencaga, A. Albayrak, and J. Wei, "Towards identification of relevant variables in the observed aerosol optical depth bias between modis and aeronet observations," in *AIP Conference Proceedings*, Vol. 1553 (American Institute of Physics, 2013) pp. 69–76.
21. M. I. Jordan and T. M. Mitchell, "Machine learning: Trends, perspectives, and prospects," Science **349**, 255–260 (2015).
22. T. Hastie, R. Tibshirani, J. H. Friedman, and J. H. Friedman, *The elements of statistical learning: data mining, inference, and prediction*, Vol. 2 (Springer, 2009).
23. S. Manabe and R. T. Wetherald, "Thermal equilibrium of the atmosphere with a given distribution of relative humidity," (1967).
24. J. Kiehl and B. Briegleb, "The relative roles of sulfate aerosols and greenhouse gases in climate forcing," Science **260**, 311–314 (1993).