



# Using Machine Learning Techniques to Improve the Performance of Numerical Weather Prediction Models

O. Sampath, Yaramala Venkata Dharani\*

Department of CSE, Rajeev Gandhi Memorial College of Engineering & Technology

Nandyal .India

reddyydharani@gmail.com

**Abstract:** The the security of industrial supply Chains (ISCs) has progressed with the incorporation of industrial internet of things (IIoT) and Blockchain (BC) technology, presenting sturdy defense in opposition to cyber attacks and ensuring operational resilience. This work examines lightweight machine learning algorithms for real-time cyber-attack detection using the WUSTL-IIOT-2021 dataset to enhance ISC safety. feature choice techniques, which include Mutual facts (MI) and further timber (ET), have been utilized to figure the most pertinent features, thereby diminishing computational complexity while retaining efficiency. This study offers a evaluation technique for assessing machine learning models, emphasizing their efficacy in figuring out cyber-attacks in a blockchain-enabled data security Context. The consequences indicate that the voting Classifier attained premiere overall performance, achieving a flawless accuracy of one hundred% with MI-decided on features and ninety nine% accuracy with ET-selected capabilities, highlighting its talent in particular and dependable chance detection. those findings underscore the importance of customized function selection and streamlined algorithms in improving cybersecurity for IIoT and blockchain-enabled records safety systems, facilitating efficient and scalable real-time applications.

**Keywords:** Web Server, Web Database, Service Provider, Remote User, Data Processing, User Queries, Dataset Storage, Weather Prediction Accuracy, Result,s Prediction Type, Ratio Train & Test Data, Bar Chart, Visualization, Profile Management, Register & Login Data, Retrieval Data Storage”.

## 1. INTRODUCTION

Improvements in processing power have elevated the evolution of “numerical weather prediction (NWP)” models, facilitating widespread numerical calculations for meteorological forecasting. The “Korea Meteorological administration (KMA)” has applied the “global Seasonal forecasting device version 6 (GloSea6) from the United Kingdom Met workplace for operational meteorological forecasting”.GloSea6 incorporates several components: the “atmospheric model (UM)”, land surface model “(JULES), ocean model (NEMO), and sea ice model (CICE)”.

The modeling procedure commences with preprocessing, throughout which is divided into grids, and for each grid they collect initial and auxiliary data, referred to as analytical fields. those fields are in the end employed to generate enter for the forecasting version, taking off numerical calculations.Due of the considerable

computational necessities of “GloSea6, the KMA presents a low-decision version, Low GloSea6, designed” for researchers missing supercomputing capabilities.

Although Low GloSea6 has a dwindled resolution, it although demands great computational sources and demonstrates significant enter/output (I/O) operations, hence requiring green I/O optimization. Researchers, mainly in atmospheric sciences, can also recollect manual performance tweaking through trial and error to be inefficient. Introducing machine -based machine learning methodology that optimizes hardware and software parameters inside low GLOSEA6.

This article offers a power optimization technique for low GLOSEA6, using machine learning and benchmarking counter. Especially we: Establish an intensive method for overall performance move-validation and empirically validate it. Classify critical statistics for the course of the Hardware Platform Parameters for GloSea6 Internal Software Parameters, extracting enormous parameters via model and facts validation. Hire Darshan, a “light-weight I/O profiling tool, to accumulate complete I/O characteristics” and validate outcomes the usage of runtime records for I/O overall performance pass-verification. Showcase the relevance of various machine learning methodologies to clarify complex correlations among execution hardware and low GloSea6 inner parameters, permitting overall performance cross-inference on novel hardware platforms. Expand the proposed approach across the workflow, positioning it as a versatile technique relevant beyond Low GloSea6.

This “paper is organized as follows: section II examines pertinent research; section III delineates GloSea6 and the profiling instrument applied for overall performance records acquisition”; phase IV elucidates the hardware and software program optimization method, encompassing the dataset and version applied Section in analyzes experiments completed using the optimization methodology alongside the model and facts validation; and section VI articulates conclusions and prospective avenues for exploration.

## 2. RELATED WORK

The progression of meteorological forecasting systems has been characterized by brilliant advancements, transferring from traditional “numerical weather prediction (NWP)” models to trendy “artificial intelligence (AI)-based methodologies”. conventional numerical weather prediction models, exemplified by using the ones created by using the “European Centre for Medium-range climate Forecasts (ECMWF)”, depend on the decision of tricky mathematical equations that dictate atmospheric dynamics. those fashions, despite the fact that fundamental, often need tremendous computational sources and may show off constraints in exactly forecasting sure climate phenomena.

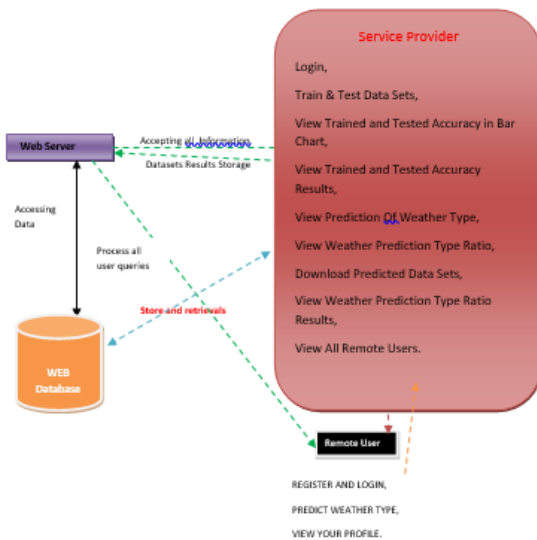
Recent improvements have included “AI and machine learning methodologies” to improve forecasting precision and efficacy. “Google's DeepMind has unveiled GraphCast, an AI version” which could produce global climate forecasts with extensive accuracy up to ten days in advance, functioning extra swiftly and economically than conventional methods. The ECMWF has evolved an AI machine that complements excessive-accuracy forecasts to 15 days earlier, surpassing conventional techniques by means of more or less 20% in critical parameters. furthermore, models consisting of Pangu-climate have illustrated the skills of AI in meteorological predictions. Pangu-weather employs deep studying to generate excessive-resolution worldwide forecasts,

exceeding conventional NWP fashions in each velocity and precision. furthermore, Google's SEEDS makes use of diffusion fashions to assess uncertainty in weather forecasting, ensuing in a massive lower in computational charges even as enhancing the depiction of severe climate phenomena.

those achievements spotlight a paradigm change in meteorology, wherein AI-pushed fashions beautify and, in positive times, exceed traditional forecasting techniques. The incorporation of AI expedites forecasting and improves prediction accuracy, specially for intense climate occurrences, consequently aiding in better disaster preparedness and reaction approaches.

### 3. MATERIALS AND METHODS

The suggested system is an all-encompassing weather forecast platform which include major modules: Service Company and far flung user. The service issuer module lets in authorized humans to effectively manage and examine meteorological data. Crucial capabilities embody training and testing datasets, displaying accuracy via bar charts, inspecting distinctive prediction consequences, studying weather kind ratios, downloading anticipated datasets, and tracking registered distant users. This module ensures particular and effective climate forecasting via the usage of state-of-the-art data processing methodologies. The far flung person module enables people to sign up and receive custom designed services. After a success registration and authentication, customers might also log in to forecast weather conditions and get admission to their money owed. This consumer-targeted method guarantees accessibility and interplay, handing over real-time weather forecasts customized to personal requirements. The incorporation of those modules permits smooth interplay between data resources and end-customers, enhancing the general efficacy of the weather prediction system.



“Fig.1 Proposed Architecture”

Figure 1 depicts the interplay amongst a web server, a web database, a service provider, and a remote user. The web server handles all person inquiries and communicates with the web database to store and retrieve facts. The service provider is capable of training

datasets, analyzing accuracy results, forecasting weather kinds, and administering customers. Remote users are capable of register, log in, forecast weather situations, and access their profiles. Information traverses among additives thru dotted traces, with various colors signifying tactics such as data retrieval, storage, and processing. The device allows uninterrupted conversation some of the database, server, and users for meteorological forecasting services.

### i) Service Provider

In this module, the service provider must log in using the valid user name and password. After a successful login, he can perform sports such as training and testing units. look at trained and tested Accuracy through Bar Chart, assessment trained and tested Accuracy outcomes, analyze weather type Predictions, verify weather Prediction type Ratios, download predicted data sets, compare weather Prediction type Ratio results, View All remote users.

### ii) View and Authorize Users

In this module, the administrator can view the listing of all registered customers. The administrator may additionally access user details, including username, e mail, and address, and has the authority to authorize people.

### iii) Remote User

This module has n users. users need to register prior to performing any operations. After registration, the user's information will be recorded in the database. After successful registration, he must log in using the permitted username and password. Upon successful login, the user will perform actions such as“REGISTER AND LOGIN, PREDICT WEATHER TYPE, VIEW YOUR PROFILE”.

### iv) Algorithms:

A **decision Tree classifier** is a supervised learning technique that divides data into subsets consistent with characteristic values, creating a tree-like structure. every node signifies a decision rule, while leaves denote magnificence labels. It adeptly manages each numerical and specific facts, despite the fact that might also overfit within the absence of pruning processes. decision trees are drastically hired in diverse domains, together with clinical prognosis and fraud detection ([17], [18]). Their readability and ease render them a favored alternative for categorization jobs.

```

GenDecTree(Sample S, Features F)
Steps:
1. Ifstopping_condition(S, F) = true then
   a. Leaf = createNode()
   b. leafLabel = classify(s)
   c. return leaf
2. root = createNode()
3. root.test_condition = findBestSpilt(S,F)
4. V = {v | v a possible outcomeofroot.test_condition}
5. For each value v ∈ V:
   a. Sv = {s | root.test_condition(s) = v and s ∈ S };
   b. Child = TreeGrowth (Sv, F);
   c. Add child as descent of root and label the edge {root → child} as v
6. return root

```

Fig.2 Decision Tree Pseudocode

**Gradient Boosting:**Increasing the gradient is access to learning a file, which gradually creates several weak students, usually decisive trees to limit prediction errors. It enhances a loss characteristic by gradient descent, enhancing model precision with each iteration. “Gradient Boosting methodologies, such XGBoost and LightGBM”, offer advanced overall performance in established data tasks in Kaggle contests ([20]).notwithstanding their computational needs, they supply strong performance via well handling missing values and interactions, rendering them appropriate for fraud detection, advice structures, and risk evaluation.

**“K-Nearest Neighbors (KNN) is a non-parametric classification technique”** that allocates class labels consistent with most people vote of the k-nearest data points within the feature space. It is straightforward but efficacious for pattern reputation and anomaly detection. KNN depends on distance measures such as Euclidean or New York distance, rendering it vulnerable to characteristic scaling. Although KNN is effective with small datasets, it encounters difficulties with excessive-dimensional records and huge-scale applications due to its great processing complexity.

$$"y = \arg \max \sum_{i \in K} I(y_i = c) (2) "$$

---

The pseudocode of classical KNN

---

**Input:** X: training data, Y: class labels of X, K: number of nearest neighbors.  
**Output:** Class of a test sample x.

**Start**  
 Classify (X,Y,x)  
 1. *for* each sample x *do*  
     Calculate the distance:  $d(x, X) = \sqrt{\sum_{i=1}^n (x_i - X_i)^2}$   
     *end for*  
 2. Classify x in the majority class:  $C(x_i) = \operatorname{argmax}_k \sum_{X_j \in KNN} C(X_j, Y_k)$   
**End**

---

Fig.3 KNN Pseudocode

**Logistic Regression** is a statistical version hired for binary type duties. The logistic function is utilized to convert enter statistics into chances, facilitating decision-making in line with a particular threshold. The method is notably applied in clinical diagnosis, credit score evaluation, and spam identification ([18]). Although Logistic Regression is simple and interpretable, it presupposes linear relationship between independent elements and logo-dependent variables, potentially constraining its efficacy on difficult, non-linear problems with out function engineering.

$$"P(y = 1|X) = \frac{1}{1 + e^{-(w^T X + b)}} (3) "$$

---

1: **Input:** Training data  
 2: **Begin**  
 3: For i = 1 to k  
 4: For each training data instance  $d_i$ .  
 5: Set the target value for the regression to  $z_i = \frac{y_i - P(1|d_j)}{[P(1|d_j)(1 - P(1|d_j))]}$   
 6: Initialize the weight of instance  $d_j$  to  $[P(1|d_j)(1 - P(1|d_j))]$   
 7: Finalize a  $f(j)$  to the data with class value ( $Z_j$ ) and weight ( $w_j$ )  
 8: **Classical label decision**  
 9: Assign (class label: 1) if  $P_{id} > 0.5$ , otherwise (class label: 2)  
 10: **End**

---

Fig.4 Logistic Regression Pseudocode

“**Naïve Bayes** is a probabilistic classifier” derived on Bayes' theorem, which presumes independence among features. notwithstanding this strong assumption, it excels in text class, spam filtering, and sentiment evaluation ([19]). Naïve Bayes fashions show off computational performance and necessitate minimal training information. editions incorporate “Gaussian, Multinomial, and Bernoulli Naïve Bayes, each tailor-made” for distinct information distributions. Naïve Bayes may additionally stumble upon difficulties when feature dependencies are present, adversely affecting accuracy in practical packages.

$$"y = \arg \max P(y = c) \prod_{i=1}^n P(x_i | y = c) (4) "$$

```

1. for q = 1 ... s // loop for each mining model's element
2.   μ[q] = 0; // initialization of mining model's elements
3. for j = 1 ... μ // loop for each vector
4.   μ[d[j],p]++; //increment count of vectors for value xj,p of vector xj;
5.   for k = 1 ... p-1 // loop for each attribute
6.     μ[φ(k-1)+(d[j, k]-1)-φ(0)+ d[j, p]]++; // increment count of vectors with
// value xj,k and value xj,p
7.   end for;
8. end for;

```

Fig.5 Naïve Bayes Pseudocode

**Random Forest** is a technique of learning a file that generates numerous decision - making trees and consolidates their predictions to increase accuracy and durability. It mitigates overfitting the usage of bootstrap sampling and random function selection ([17]). “Random forest is appreciably utilized in finance, healthcare, and bioinformatics” due to its robust generalization functionality. it could efficiently control absent values and erratic records. nevertheless, it may incur big computational costs, mainly with extensive trees and expansive characteristic regions.

“**Support Vector Machines (SVM)**” are robust classification models that delineate data through hyperplanes in a excessive-dimensional space. “Support Vector machine (SVM)” maximizes the margin between instructions to improve generalization and resilience. It facilitates each linear and non-linear classification through kernel functions, including “radial foundation function (RBF) and polynomial kernels ([18])”. “Support Vector Machines (SVM)” are proficient in image popularity, textual content classification, and bioinformatics; yet, they may require big processing sources, specifically with considerable datasets featuring problematic barriers.

$$"f(X) = \text{sign} \left( \sum_{i=1}^m \alpha_i y_i K(x_i, X) + b \right) (5) "$$

Training Model for SVM

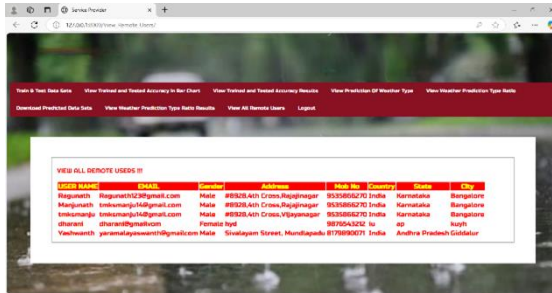
```

Input: D=[X,Y]; X(array of input with m features), Y(array of class labels)
Y=array(C) // Class label
Output: Find the performance of the system
function train_svm(X,Y, number_of_runs)
  initialize: learning_rate=Math.random();
  for learning_rate in number_of_runs
    error=0;
    for i in X
      if (Y[i] * X[i]^w) < 1 then
        update : w=w + learning_rate * ((X[i]*Y[i])*(-2*(1/number_of_runs)*w)
      else
        update: w=w+learning_rate * (-2*(1/number_of_runs)*w)
      end if
    end
  end
end

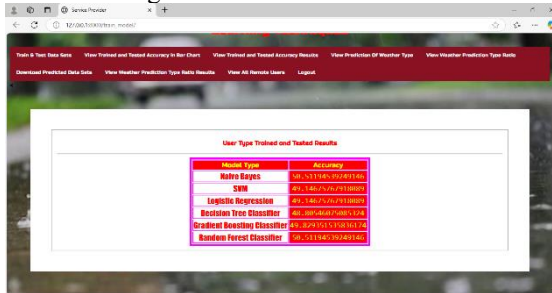
```

Fig.6 SVM Pseudocode

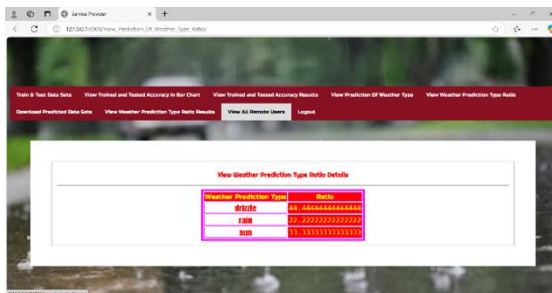
### 4. RESULTS & DISCUSSION



“Fig.7 View All Remote Users”



“Fig.8 Algorithms Accuracy”



“Fig.9 Weather Prediction Ratio”

### 5. CONCLUSION

This study offered access to machine learning for optimizing the hardware and software aspects of medical products. Low Glosea6, a medical weather forecasting software, was chosen, and a data set was constructed that included the internal tool settings, hardware platform specs, and metrics resulting from their combinations. Before using the machine learning model, the data file was shown, and the integrity of the regression version, trained on a tiny data file, was checked using the "Cross-Validation (LOOCV)" approach. Trained machine learning was used to discover the optimal hardware platform requirements and internal parameters for low GLOSEA6 in new research conditions, which matched the actual combination parameters. The projected execution time determined from the combination of factors revealed 16% of the errors

when compared to the actual execution time, demonstrating the effectiveness of access at the implementation prognosis time.

The suggested optimization method can be utilized to improve the efficacy of additional “high-performance computing (HPC)” research programs. Similarly to weather and climate modeling, those applications encompass “computational fluid dynamics (CFD) simulations, molecular dynamics (MD) simulations, and quantum chemistry calculations”. Researchers utilizing “high-performance computing (HPC)” applications frequently depend on supercomputing middle personnel for utility optimization, and our method can accelerate this manual performance enhancement process.

Two avenues for future research are delineated concerning information accessibility. To begin with, augmenting the entire quantity of data is essential. The prediction of execution time in this research was impeded by the exclusion of some hardware platform data. Consequently, accumulating supplementary “hardware/software program parameters and I/O overall performance metrics”, as recommended in prior research on improving I/O overall performance in HPC systems, would increase version efficacy. Secondly, the implementation of a benchmark-primarily based go-inference optimization technique, as originally defined in this paper, could be effective. This technique might expedite data accumulating and facilitate the acquisition of “parameter values not collected on this look at via using exchange parameters”, consequently enhancing version performance and expanding its applicability.

## REFERENCES

- [1] Concept of a Numerical Forecast Model. Accessed: Aug. 10, 2023. [Online]. Available: [http://web.kma.go.kr/aboutkma/intro/supercom/model/model\\_concept.jsp](http://web.kma.go.kr/aboutkma/intro/supercom/model/model_concept.jsp)
- [2] P. Davis, C. Ruth, A. A. Scaife, and J. Kettleborough, “A large ensemble seasonal forecasting system: GloSea6,” Dec. 2020, vol. 2020.
- [3] M. Howison, Q. Koziol, D. Knaak, J. Mainzer, and J. Shalf, “Tuning HDF5 for Lustre file systems,” Lawrence Berkeley Nat. Lab., Berkeley, CA, USA, Tech. Rep. LBNL-4803E, 2010.
- [4] B. Behzad et al., “Taming parallel I/O complexity with auto-tuning,” in Proc. Int. Conf. High Perform. Comput., Netw., Storage Anal., 2013, p. 68.
- [5] B. Behzad, S. Byna, Prabhat, and M. Snir, “Optimizing I/O performance of HPC applications with autotuning,” ACM Trans. Parallel Comput., vol. 5, no. 4, pp. 1–27, Mar. 2019, doi: 10.1145/3309205.
- [6] S. Robert, S. Zertal, and G. Goret, “Auto-tuning of IO accelerators using black-box optimization,” in Proc. Int. Conf. High Perform. Comput. Simulation (HPCS), Jul. 2019, pp. 1022–1027, doi: 10.1109/HPCS48598.2019.9188173.
- [7] A. Bağbaba, X. Wang, C. Niethammer, and J. Gracia, “Improving the I/O performance of applications with predictive modeling based auto-tuning,” in Proc. Int. Conf. Eng. Emerg. Technol. (ICEET), Oct. 2021, pp. 1–6, doi: 10.1109/ICEET53442.2021.9659711.

- [8] S. Valcke and R. Redler, “The OASIS coupler,” in *Earth System Modelling*, vol. 3. Berlin, Germany: Springer, 2012, pp. 23–32, doi: 10.1007/978-3-642-23360-9\_4.
- [9] Analysing UM Outputs. Accessed: Feb. 14, 2023. [Online]. Available: [http://climate-cms.wikis.unsw.edu.au/Analysing\\_UM\\_outputs](http://climate-cms.wikis.unsw.edu.au/Analysing_UM_outputs)
- [10] Unidata | NetCDF. Accessed: Nov. 28, 2022. [Online]. Available: <https://www.unidata.ucar.edu/software/netcdf/>
- [11] Icons8. Free Icons, Clipart Illustrations, Photos, and Music. Accessed: Jul. 18, 2023. [Online]. Available: <https://icons8.com>
- [12] P. Carns, R. Latham, R. Ross, K. Iskra, S. Lang, and K. Riley, “24/7 characterization of petascale I/O workloads,” in *Proc. 2009 Workshop Interfaces Archit. Sci. Data Storage*, Sep. 2009, pp. 1–10.
- [13] Darshan Introduction. Accessed: Aug. 10, 2023. [Online]. Available: <https://wordpress.cels.anl.gov/darshan/wp-content/uploads/sites/54/2014/08/iiswc-2014-darshan-instrumentation.pdf>
- [14] R. Ross, D. Nurmi, A. Cheng, and M. Zingale, “A case study in application I/O on Linux clusters,” in *Proc. ACM/IEEE Conf. Supercomput.*, New York, NY, USA, Nov. 2001, p. 11, doi: 10.1145/582034.582045.
- [15] S. Herbein, D. H. Ahn, D. Lipari, T. R. W. Scogland, M. Stearman, M. Grondona, J. Garlick, B. Springmeyer, and M. Taufer, “Scalable I/O aware job scheduling for burst buffer enabled HPC clusters,” in *Proc. 25th ACM Int. Symp. High-Perform. Parallel Distrib. Comput.*, New York, NY, USA, May 2016, pp. 69–80, doi: 10.1145/2907294.2907316.
- [16] R Package. Accessed: Apr. 26, 2023. [Online]. Available: <https://cran.rproject.org/src/contrib/Archive/>
- [17] R. Genuer, J.-M. Poggi, and C. Tuleau, “Random forests: Some methodological insights,” 2008, arXiv:0811.3619.
- [18] T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning* (Springer Series in Statistics). New York, NY, USA: Springer, 2009, doi: 10.1007/978-0-387-84858-7.
- [19] Package RandomForest. Accessed: Apr. 26, 2023. [Online]. Available: <https://cran.r-project.org/web/packages/randomForest/randomForest.pdf>
- [20] PackageGBM. Accessed: Apr. 26, 2023. [Online]. Available: <https://cran.rproject.org/web/packages/gbm/gbm.pdf>
- [21] Tekie. (Sep. 23, 2022). SSD vs HDD—Comparing Speed, Lifespan, Reliability. Accessed: Feb. 15, 2023. [Online]. Available: <https://tekie.com/blog/hardware/ssd-vs-hdd-speed-lifespan-and-reliability/>
- [22] L. M. Rea and R. A. Parker, *Designing & Conducting Survey Research A Comprehensive Guide*, 3rd ed. San Francisco, CA, USA: Jossey-Bass, 2012.

[23] R. Kabacoff, *R in Action: Data Analysis and Graphics with R and Tidyverse*, 3rd ed. New York, NY, USA: Simon and Schuster, 2022. *Annu. Ubiquitous Comput., Electron. Mobile Commun. Conf. (UEMCON)*, Dec. 2021, pp. 313–318.

**Open Access** This chapter is licensed under the terms of the Creative Commons Attribution-NonCommercial 4.0 International License (<http://creativecommons.org/licenses/by-nc/4.0/>), which permits any noncommercial use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license and indicate if changes were made.

The images or other third party material in this chapter are included in the chapter's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the chapter's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder.

