ALBERT M.W. YAU

GitHub: mwyau
Email: albert@mwyau.com
LinkedIn: albertmwyau
Google Scholar: V-2nZZ8AAAAJ

Data Engineer with 5 years of experience in data and machine learning infrastructure, contributing to the growth of a successful fintech start-up. Previously a Climate Scientist with over 10 years of statistics, machine learning, Linux, and High-Performance Computing (HPC) research experience, and an entrepreneur in the early 2000s.

WORK EXPERIENCE

Data Engineer, Data Platform, dv01 / Fitch Solutions, New York, NY Aug 2018 – Mar 2023

As one of the longest tenured engineers, I managed the end-to-end ETL data infrastructure and backend services:

- Spearheaded the data pipeline migration to Spark 3 as an advocate and the technical lead across all teams in the company, preventing *end-of-life* of old Google Cloud Dataproc images that would interrupt all production *ETL* jobs:
  - Upgraded the entire monolithic *Scala* codebase from 2.11 to 2.12.
  - Upgraded *Apache Spark* from 2.4 to 3.0 and upgraded all required dependencies, which is a fragile process.
  - Migrated data warehouse metadata from self-hosted *Apache Hive Metastore* to *Google Dataproc Metastore*.
  - Worked closely with data scientists to upgrade all production *R* Docker images to use the new data warehouse.
  - Managed Jira and reviewed PRs; patched an incompatibility found on the deploy day and prevented a rollback.
- Acted as the technical lead of the migration of the data warehouse from *Microsoft Azure* to *Google Cloud* at the *petabyte* scale; executed a seamless transition with no data loss and zero downtime.
- Proposed the first continuous version update strategy of the data pipeline codebase; implemented automatic update alerts and scheduled vulnerability checks through *GitHub Actions* and *Dependabot*.
- Put forward and executed the first *disaster recovery plan* in the company; scheduled routine backups of the *Apache Parquet* data warehouse and verified the restoration process to protect business-critical data.
- Improved data warehouse performance ten-fold through migration from Microsoft SQL Server to *Google BigQuery*.
- Maintained 1000s of active DAGs in *Airflow*; migrated Airflow from Compute Engine to *Google Cloud Composer*.
- Created, maintained, and migrated CI/CD pipelines in *Jenkins*, *Argo CI/CD*, *GitHub Actions* and *Terraform*.
- Created and maintained backend services in *Scala* and *Python*, including automatic file scrapers and PDF parsers, *Airflow* plugins and operators, and API endpoints using *FastAPI/SQLAlchemy/PostgreSQL*.
- Worked in an agile environment, such as performing code reviews, sprint planning, and participating in hack days.
- Supported the machine learning pipeline in *Databricks* and *Airflow* utilizing the *H2O Sparkling Water* library.

Graduate Student Researcher, SoMAS / Stony Brook University, Stony Brook, NY Jun 2012 – Aug 2018

As a PhD candidate, my expertise is in time series analysis; I studied storm tracks using very large climate model datasets:

- Administered workstations and deployed new storage systems with hundreds of terabytes of capacity.
- Aggregated, processed, and managed a century's worth of climate data from over 30 global climate models in the CMIP5 archive, originating from 20 international institutions, on local storage systems. This streamlined access and analysis for researchers, contributing to multiple highly cited publications, with two cited in the *Sixth Assessment Report* (AR6) of the United Nations *Intergovernmental Panel on Climate Change* (IPCC).
- For my dissertation, I developed statistical frameworks based on principal component analysis (PCA) and canonical correlation analysis (CCA); the work was published in two parts in the *Journal of Climate* in 2020 and 2022.
- Investigated using *deep learning* (Convolutional Neural Network) to predict extreme weather events.
- Departed the position to join an early-stage Series A fintech start-up, *dv01*.

Oceanographer, Coastal Environments, La Jolla, CA Mar 2010 – Jul 2011

As a coastal engineering consultant, I performed data collection, analysis, and modeling:

- Collected seabed sediment samples in the San Diego Harbor; gathered water flow data in San Dieguito Lagoon using GPS-equipped drifters.
- Modeled beach erosion process, tidal dynamics, and tsunami waves caused by distant earthquakes.
- Discovered short period sea-level fluctuations in the Santa Barbara Channel.

HPC Support Assistant, The Chinese University of Hong Kong, Hong Kong 2005 – 2008

I worked in the High-Performance Computing (HPC) Support Team under the Information Technology Services Centre to support the operations of on-campus supercomputing clusters:

- Assisted users including graduate students, postdoctoral researchers, and faculty members in using the HPC clusters:
  - Gathered requirements and installed dependent libraries.
  - Compiled, fine-tuned, and deployed scientific computing programs.
  - Debugged and troubleshooted programs on the HPC clusters, working closely with research users.
- Conducted *User Acceptance Testing* (UAT) of a newly constructed HPC cluster.

Founder, XDDD.org 2002 – 2005

I operated web hosting and webmail services in the early 2000s:

- Built a server which was hosted in the HKNet data center in Hong Kong.
- Implemented *LAMP* (Linux/Apache/MySQL/PHP) stack on RedHat Linux 9.
- Managed user experience, provided technical support, and created dashboards; responded to security incidents.
- Offered competitive pricing at HKD$100/100MB/year per user, supporting up to 300 users with a 40GB hard drive.

SKILLS

Programming: Python (NumPy/Pandas/PyTorch/FastAPI), Scala, Spark, SQL, RESTful API, Linux shell scripts.

Cloud Technologies: Google Cloud Platform, Microsoft Azure, Databricks, Kubernetes, Docker, Cloudflare.

Data Engineering/DevOps: ETL, Airflow, Hive, Google BigQuery/Cloud Run/Functions, GitHub Actions, Terraform.

Online Course Certifications: Deep Neural Networks with PyTorch, Natural Language Processing with Attention Models, Machine Learning Engineering for Production (MLOps) Specialization, Deep Learning Specialization, Applied Machine Learning in Python Specialization, Data Engineering on Google Cloud Platform Specialization.

EDUCATION

MPhil ITPA / SoMAS, Stony Brook University, *Marine and Atmospheric Science* 2020
Dissertation: "Finding Storm Track Activity Metrics That Are Highly Correlated with Weather Impacts"
Advisor: Prof. Edmund K.M. Chang

MS Scripps Institution of Oceanography, UC San Diego, *Oceanography* 2009
Advisor: Prof. Myrl C. Hendershott

BSc The Chinese University of Hong Kong, *Theoretical Physics* 2008
Double minored in *Computer Science* and *Mathematics*

HONORS AND AWARDS

Yasumoto International Exchange Scholarship 2006
Exchange Student at University of Toronto

Chung Chi College Summer Study Abroad Program 2005
Exchange Student at UC Berkeley

CN Yang Scholarship 2005
Department of Physics, The Chinese University of Hong Kong

RESEARCH EXPERIENCE

Dissertation, Stony Brook University, Stony Brook, NY 2020

Advisor: Prof. Edmund K.M. Chang

Dissertation published in the *Journal of Climate*:

- "Finding Storm Track Activity Metrics That Are Highly Correlated with Weather Impacts:"
  - "Part I: Frameworks for Evaluation and Accumulated Track Activity."
  - "Part II: Estimating Precipitation Change Associated with Projected Storm Track Change over Europe."

Summer Internships in Parallel Computational Science, NCAR, Boulder, CO 2015
Data Science Intern, Application Scalability and Performance Group

Advisor: Dr. Kevin Paul and John Dennis

Starting from scratch, designed and created a parallel cyclone tracker in Python in 10 weeks, and successfully benchmarked on NCAR's Yellowstone supercomputer:

- "PyStormTracker: A Parallel Object-Oriented Cyclone Tracker in Python."

Summer Undergraduate Research Exchange, Caltech, Pasadena, CA 2007
Visiting Student Researcher, Division of Geological and Planetary Sciences

Advisor: Prof. Yuk L. Yung

Undergraduate research that was presented as a poster in AGU Fall Meeting 2007.

- "Solar-cycle response in global climate models assessed by IPCC AR4."

Hong Kong Observatory Summer Internship Program, HKO, Hong Kong 2006
Summer Intern, Aviation Weather Forecast and Warning Services

Advisor: Dr. Ping-Wah Li

- "Ingesting of data from a Doppler LIDAR (LIght Detection And Ranging) into the Local Analysis and Prediction System."

TEACHING EXPERIENCE

Stony Brook University, Stony Brook, NY Aug 2011 – May 2013
Teaching Assistant / Guest Lecturer, School of Marine and Atmospheric Sciences

- ATM102 Weather and Climate
- ATM201 Introduction to Climate and Climate Change
- ATM397 Air Pollution and Its Control
- ATM320 Spatial Data Analysis using Matlab (Guest Lecturer)

PUBLICATIONS

Journal Publications

- Chang, E.K.M., Yau, A.M.W. & Zhang, R. (2022), "Finding Storm Track Activity Metrics That Are Highly Correlated with Weather Impacts. Part II: Estimating Precipitation Change Associated with Projected Storm Track Change over Europe." *Journal of Climate*, **35**, 2423-2440. https://doi.org/10.1175/JCLI-D-21-0259.1

- Yau, A.M.W. & Chang, E.K.M. (2020), "Finding Storm Track Activity Metrics That Are Highly Correlated with Weather Impacts. Part I: Frameworks for Evaluation and Accumulated Track Activity." *Journal of Climate*, **33**, 10169-10186. https://doi.org/10.1175/JCLI-D-20-0393.1

- Chang, E.K.M. & Yau, A.M.W. (2016), "Northern Hemisphere winter storm track trends since 1959 derived from multiple reanalysis datasets." *Climate Dynamics*, **47**, 1435-1454. https://doi.org/10.1007/s00382-015-2911-8

- Chang, E.K.M., Ma, C.G., Zheng, C. & Yau, A.M.W. (2016), "Observed and projected decrease in Northern Hemisphere extratropical cyclone activity in summer and its impacts on maximum temperature." *Geophysical Research Letters*, **43**, 2200-2208. https://doi.org/10.1002/2016GL068172

- Chang, E.K.M., Zheng, C., Lanigan, P., Yau, A.M.W. & Neelin, J.D. (2015), "Significant modulation of variability and projected change in California winter precipitation by extratropical cyclone activity." *Geophysical Research Letters*, **42**, 5983-5991. https://doi.org/10.1002/2015GL064424

Unpublished Manuscripts

- Yau A.M.W. & Hendershott M.C., "Short Period Sea-level Fluctuations in the Santa Barbara Channel-Santa Maria Basin."

PRESENTATIONS

Poster Presentations

American Meteorological Society Annual Meeting

- **99th AMS Annual Meeting**, Jan 2019, Phoenix, AZ, "Quantifying Storm-Track Variability and Impacts Using Accumulated Cyclone Track Activity."
- **96th AMS Annual Meeting**, Jan 2016, New Orleans, LA, "PyStormTracker: A Parallel Object-Oriented Cyclone Tracker in Python."

American Geophysical Union Fall Meeting

- **AGU Fall Meeting 2013**, Dec 2013, San Francisco, CA, "Impacts of Background Field Removal on the Seasonal Cycle and Trend of Cyclone Statistics."
- **AGU Fall Meeting 2007**, Dec 2007, San Francisco, CA, "Solar-cycle response in global climate models assessed by IPCC AR4."

Workshop Presentations

Pacific Rim Application and Grid Middleware Assembly

- **PRAGMA 14**, Mar 2008, Taichung, Taiwan, "Solar Variability and Climate Change."

CONFERENCES AND WORKSHOPS

- **96th AMS Annual Meeting**, Jan 2016, New Orleans, LA
- **20th Annual CESM Workshop**, Jun 2015, Breckenridge, CO
- **AGU Fall Meeting 2013**, Dec 2013, San Francisco, CA
- **PRAGMA 18**, Mar 2010, San Diego, CA
- **CAM Tutorial 2009**, Jul 2009, Boulder, CO
- **PRAGMA 14**, Mar 2008, Taichung, Taiwan

LANGUAGES

Mandarin and Cantonese: Native language proficiency
English: Full professional proficiency

OTHER INTERESTS AND HOBBIES

- 25 years of PC building and troubleshooting experience.
- Artificial Intelligence: I used my gaming GPU to train character-level RNN and nanoGPT with Chinese texts.
- Audio: I am fascinated by the physics of sound and enjoy reading objective reviews on Audio Science Review.
- Flightaware flight tracking: Raspberry Pi running PiAware with ADS-B antennas in the attic (feeder statistics).
- Home automation: whole home solar energy monitor using IoTaWatt / PVOutput and Home Assistant; integrated Home Assistant dashboard showing personal weather station data, electrical loads, heating oil level, solar generation, water usage, and refrigerator/freezer temperature alarms.
- Home networking lab on a 10GbE Ubiquiti network stack:
    - Ryzen 9 5950X server with 128GB ECC memory and 134TB ZFS arrays; Raspberry Pi's, and Jetson Nano.
    - Redundant Pi-hole DNS servers for family and ad blocking, using multiple DoH upstream for resiliency.
- Music: I enjoy a wide variety of genres, especially baroque music; I play the recorder (SSATB), flute, harmonica (beginner), highland bagpipe (practice chanter); I passed ABRSM Grade 5 in Descant Recorder and Music Theory.
- Photography and astrophotography: I traveled to Oregon on Aug 21, 2017 to capture the total solar eclipse on video.
- Scouting: I achieved Challenger Award rank in Hong Kong; I enjoy hiking, camping, and traveling.