ЭЛЕКТРОННАЯ БАЗА МНОГОЛЕТНИХ ДАННЫХ ГЛОБАЛЬНОГО РАДИОТЕПЛОВОГО ПОЛЯ СИСТЕМЫ ОКЕАН – АТМОСФЕРА В КОНТЕКСТЕ ЗАДАЧ ИССЛЕДОВАНИЯ ВАРИАЦИЙ КЛИМАТА ПЛАНЕТЫ И АТМОСФЕРНЫХ КАТАСТРОФ

Д.М. Ермаков[1], М.Д. Раев[2], А.И. Суслов[2], Е.А. Шарков[2]

[1] *Институт радиотехники и электроники РАН, Фрязинское отделение,*
141120 Фрязино, пл. Введенского, 1
E-mail: dima@ire.rssi.ru ;

[2] *Институт космических исследований РАН,*
117997 Москва, Профсоюзная, 84/32
E-mail: mraev@iki.rssi.ru

В работе описана развиваемая в ИКИ РАН электронная база многолетних данных измерений спутникового СВЧ радиометра SSM/I. Создание базы данных вызвано необходимостью использования информации о глобальном радиотепловом поле системы океан – атмосфера в задачах исследования климата Земли и, в перспективе, в ряде других актуальных задач, решаемых в ИКИ РАН. Сформулирован оригинальный принцип интерпретации спутниковых данных, ставший идеологической основой создаваемой базы данных. Описано разработанное и внедренное к настоящему моменту специальное программное обеспечение. Очерчены ближайшие перспективы развития создаваемой базы данных.

Введение

Одним из важнейших климатообразующих факторов на Земле считается многомасштабное (в пространстве и времени) взаимодействие океана и атмосферы, складывающееся из многообразных процессов обмена энергией, импульсом и веществом. Основным средством получения мгновенных характеристик этого взаимодействия (температуры атмосферы и поверхности океана, скорости приповерхностного ветра, общего влагосодержания атмосферы, интенсивности осадков и т. д.) в глобальных масштабах является спутниковая СВЧ радиометрия. Из функционирующих в настоящее время спутниковых СВЧ радиометров следует особо выделить приборы SSM/I, популярные в научном сообществе благодаря беспрецедентной надежности измерений и долгого времени стабильной работы на орбите в рамках американского проекта Defense Meteorological Satellite Program, DMSP (http://dmsp.ngdc.noaa.gov/dmsp.html).

SSM/I — семиканальный радиометр, принимающий линейно поляризованное излучение на частотах 19,35; 22,235; 37,0 и 85,5 ГГц. На всех частотах, кроме 22,235 ГГц, измеряется как горизонтально, так и вертикально поляризованное излучение; на 22,235 ГГц — только вертикально поляризованное. Пространственный шаг измерений на поверхности Земли равен 12,5 км для каналов 85,5 ГГц и 25 км для других каналов (при разных размерах пятна разрешения). Полоса обзора составляет около 1400 км в ширину, геометрия сканирования — коническая. SSM/I базируется на американских спутниках серии DMSP, измерения практически полностью покрывают поверхность Земли. (см., например, http://podaac.jpl.nasa.gov:2031/SENSOR_DOCS/ssmi.html).

Ряд электронных архивов (в основном, в США) специализируется на хранении данных SSM/I. Ознакомление с имеющимися данными можно осуществлять через информационные порталы этих архивов в сети Internet, члены научного сообщества имеют возможность заказать необходимую информацию в режиме on-line. Авторами был установлен надежный контакт с Глобальным Гидрологическим Исследовательским Центром США, GHRC (http://ghrc.msfc.nasa.gov), являющимся одним из основных хранителей спутниковых данных с аппаратов серии DMSP. В результате были получены данные измерений SSM/I за следующие периоды:

Год	***Месяцы***	***Приборы***
1995	Май – декабрь	F10 и F13
1996	Январь – декабрь	F10 и F13
1997	Январь – декабрь	F13, F14 и F15
1998	Январь – декабрь	F13, F14 и F15
1999	Январь – декабрь	F13, F14 и F15
2000	Январь – декабрь	F13, F14 и F15
2001	Январь – декабрь	F13, F14 и F15
2002	Январь – декабрь	F13, F14 и F15
2003	Январь – декабрь	F13, F14 и F15

Общий объем данных в компрессированном виде к настоящему времени составляет величину порядка 150 Гбайт.

Электронная база данных SSM/I в ИКИ РАН

Значение информации о глобальном радиотепловом поле системы океан – атмосфера для многообразия работ в области климатологии и исследовании атмосферы и океана, ведущихся в ИКИ РАН, сделало актуальным создание собственной базы дистанционных данных SSM/I. В основу построения этой базы данных положен принцип рассмотрения дистанционных данных в форме последовательностей рядов наблюдений: пространственных (глобальный охват Земли с возможностью зонирования) и временных (многолетние ежедневные наблюдения отдельных зон и всего земного шара). Последовательность радиотепловых измерений рассматривается при этом не как механическое объединение данных из нескольких файлов, соответствующих последовательным моментам съемки или соседним точкам на поверхности Земли, а является, с точки зрения пользователя, основной структурной единицей базы данных, длина которой определяется решаемыми задачами и операциями обработки. Конкретные характеристики рядов данных (источники данных, пространственная и временная протяженность, дискретизация, осреднение и т. д.) определяются на основании параметров запроса пользователя, генерируемого с учетом длины пространственно-временного ряда. Выходные данные могут быть записаны в один или несколько файлов. Наиболее наглядным методом представления полученных данных является формирование серии изображений (например, в виде видеоклипа — анимационный метод), либо построение характеристик длительных временных рядов радиояркостной температуры для определенных областей земной поверхности.

Такой подход к построению базы данных, по мнению авторов, наиболее адаптирован к задачам для анализа глобальных изменений системы океан-атмосфера на сезонных, годовых и многолетних временных масштабах. В то же время, он не имеет полных аналогов среди открытых электронных архивов данных дистанционных наблюдений, которые, как правило, реализуют простейший отбор и компоновку файлов данных по введенным пользователем критериям отбора без существенной переработки структуры данных и методов представления данных в файлах. Последовательная реализация авторского подхода позволила определить наиболее адекватную структуру базы данных и формат внутреннего представления данных в ней, классифицировать основные типы данных, генерируемых по запросам пользователя; привела к созданию оригинального программного обеспечения, сочетающего требуемую гибкость и эффективность при работе пользователя с хранимыми данными.

Структура и внутреннее представление данных в базе данных

Структуру создаваемой базы данных определили требования актуальных климатологических задач, задач физического исследования процессов в системе океан-атмосфера на разных временных и пространственных масштабах. Сформулированный выше принцип получения спутниковой радиометрической информации в форме рядов данных адаптивной длины предполагает возможность оперативного доступа к динамически формируемым из общей базы фрагментам данных на жестких дисках сервера базы данных либо на компьютерах, объединенных с сервером в локальную сеть. Автономные носители (компакт-диски, магнитные ленты и т.п.), изначально рассматривавшиеся в основном с точки зрения резервного копирования, в настоящее время в связи большим объемом накопленных данных также предполагаются использовать для хранения баз данных. Таким образом, структуру базы данных можно представлять в виде дерева каталогов на различного рода носителях, которые могут включать в себя локальные и сетевые каталоги, логические диски и внешние носители. Внутри дерева каталогов данные группируются по источникам (номерам спутников), времени съемки и/или другим параметрам. Это требует создания дополнительной индексной базы данных, формируемой в процессе анализа всех физических носителей, на которых размещены файлы записи пространственно-временных последовательностей данных спутниковой радиометрической информации. Индексная база обеспечивает возможность быстрого поиска нужной области на физических носителях с использованием запросов на естественных языках и метаязыка.

Спутниковая информация, поступающая в ИКИ РАН из электронных архивов, доступных через Internet, записана в файлах формата HDF, который является общепринятым мировым стандартом хранения дистанционных данных. Развитая типология формата HDF делает естественным выбор его в качестве метаязыка (языка описания данных) создаваемой базы данных. Формат HDF был разработан в NCSA (National Center for Supercomputing Applications, http://www.ncsa.uiuc.edu/), в Университете Иллинойса, США, как стандарт организации данных для их широкого свободного обмена в научном сообществе. Основная идея, связанная с внедрением такого стандарта, состоит в том, чтобы избавить пользователей от необходимости сопровождать собственно данные пояснениями и комментариями, описывающими контекст, в котором те или иные блоки данных должны быть интерпретированы и использованы.

С этой целью в HDF введен ограниченный набор базовых типов данных. К ним относятся: числовой и строковый атрибуты, ряд (многомерный массив), растровое изображение, палитра (индексированный набор цветов), V-группа. Предполагается, что для данных произвольного происхождения можно подобрать адекватное представление в терминах одного или нескольких типов данных, определенных в HDF. Так, универсальные константы можно хранить в виде численных атрибутов, результаты эксперимента — в виде многомерного массива целых или дробных чисел, графики — в виде растровых изображений и т. д. V-группы используются для логического объединения нескольких блоков данных по аналогии с директориями файловой системы. Пользователям предоставлена возможность снабжать записанные блоки данных и файлы в целом текстовыми комментариями путем добавления строковых атрибутов.

Внутреннее представление данных в HDF подсказывает, что наиболее удобно организовать их хранение в файлах не линейно (последовательная запись поступающей информации), а иерархически. В начале каждого HDF-файла размещается заголовок, который описывает общее содержание файла и ссылается на подзаголовки отдельных типов данных. Последние содержат информацию о количестве блоков данных своего типа и ссылаются на их подзаголовки. В каждом из новых подзаголовков детализируются параметры конкретного блока данных и т. д. по иерархическому принципу до самого нижнего уровня — собственно данных. В целом, структура напоминает книжное оглавление, а роль основного текста играют сами данные.

В связи со сложностью структуры HDF-файлов рекомендованным способом для чтения и записи информации является использование стандартных библиотек. Такие библиотеки поставляются разработчиком в виде исходных кодов, адаптируемых под разные платформы и операционные системы. Авторами использовались библиотеки, откомпилированные на IBM-совместимых ПК под семейство ОС Windows в среде разработки программных продуктов MS Visual Studio 6.0.

Таким образом, внутренним стандартом представления данных в создаваемой базе данных являются файлы формата HDF. Для экономии дискового пространства к хранящейся информации применена стандартная tar- или gz-компрессия. Следует отметить, что помимо файлов, содержащих собственно данные измерений на исходной координатной сетке прибора, необходимо хранить в базе данных и информацию по географической привязке этих данных, поставляемую в виде специальных файлов по два на каждый файл данных (координаты измерений отдельно в низкочастотных и высокочастотных каналах).

Основные типы данных, генерируемых по запросу пользователя

Данные, хранящиеся в создаваемой электронной базе данных, содержат вычисленные (восстановленные по исходным спутниковым измерениям) величины радиояркостных температур во всех каналах SSM/I: на исходной сетке измерений (данные типа SWATH) и на регулярной сетке, перевод в которую осуществлен путем осреднения исходных данных в ячейках размером 0,5° широты на 0,5° долготы (данные типа GRID). Данные SWATH несут более полную и адекватную информацию, и именно их предпочтительно использовать при генерировании выходных данных, предоставляемых по запросу пользователя. Выходные данные строятся на основе данных в компрессированных HDF файлах, но имеют два принципиальных отличия от исходных данных:

1. При генерации выходных данных используются только те исходные данные, которые удовлетворяют заданным пользователем критериям отбора. Фильтрация осуществляется не только на уровне отдельных файлов, но и на уровне фрагментов информации внутри файла.
2. Выходные данные имеют принципиально более простое представление в файлах, чем исходные данные, что значительно облегчает работу с ними пользователя (в частности, не требует знания HDF).

В соответствии с принятым принципом интерпретации хранимой информации, выходные данные представляются в виде длинных рядов измерений, выполненных в фиксированных (заданных пользователем) условиях. В настоящее время такие ряды данных формируются в виде последовательности нескольких (возможно, большого количества) файлов. В перспективе рассматриваются также различные возможности синтеза одного файла более сложного типа из полученной последовательности.

Поскольку длинные ряды измерений были выполнены спутниковыми приборами за длительные интервалы времени, для их объединения в последовательность однотипных блоков данных представляется разумным перейти от измерений на собственной координатной сетке прибора (имеющей вследствие особенностей геометрии сканирования и кривизны поверхности Земли нерегулярную структуру) к регулярной сетке (с постоянным шагом по широте и долготе). Для этого используется метод осреднения данных в ячейках, аналогичный стандартному, применяемому при переходе от SWATH к GRID данным. Важной особенностью является то, что пользователь сам выбирает масштабы осреднения. При необходимости, задавшись соответствующим малым масштабом, он получает реальные данные разовых спутниковых измерений.

Итак, генерируемые выходные данные представляют собой восстановленные по измерениям SSM/I значения радиояркостных температур системы океан-атмосфера, уложенные на регулярную сетку и запи-

санные в последовательность файлов простого формата. Детальная классификация выходных данных проведена далее по двум категориям:

- условиям отбора (фильтрации) исходных данных;
- способу представления данных в файлах данных.

Условия отбора, определяемые запросом пользователя, позволяют осуществить выборку требуемых данных по следующим (независимым) критериям:

1) источник данных (отбор информации с одного из спутников серии DMSP: F8, F10, F11, F12, F13, F14, F15, ...);
2) канал наблюдений (от одного до семи измерительных каналов SSM/I);
3) интервал наблюдений (начало и конец наблюдений с точностью до года, месяца и дня);
4) зона наблюдений (глобальный охват или прямоугольник с заданными граничными широтами и долготами);
5) пространственный масштаб осреднений (шаг регулярной сетки в градусах и долях градуса, настраиваемый независимо по широте и долготе)
6) временной масштаб осреднения в единицах суток (используется для накопления и усреднения данных последовательных пролетов спутника над выбранной областью Земли).

Любые отобранные по указанным выше критериям данные в настоящий момент могут быть представлены в одном из двух видов, PCK и RAW.

Данные в формате PCK записаны в виде последовательности файлов *.pck, содержащих рассчитанные осреднением значения радиояркостных температур в узлах регулярной сетки в формате чисел с плавающей запятой двойной точности (8 байтов на число). Фактическая точность определяется максимальной точностью исходных данных (сотые доли градуса). Узлы сетки считаются упорядоченными в строки по убыванию западной и/или возрастанию восточной долготы (направление с запада на восток) при фиксированной широте, а последовательные строки — по убыванию северной и/или возрастанию южной широты (направление с севера на юг). Каждый файл *.pck содержит данные по одному интервалу временного осреднения (либо, без временного осреднения, данные за одни сутки), отфильтрованные, усредненные и объединенные с учетом прочих условий отбора: полученные от заданного источника (спутника), измеренные в выбранном канале (при выборе нескольких каналов формируется соответствующее число отдельных файлов), попавшие в заданную область поверхности Земли. Последовательность таких файлов покрывает весь выбранный пользователем временной интервал наблюдений (с ограничениями, накладываемыми возможным отсутствием необходимых исходных данных). Семантика имени pck-файла отражает описанные выше рамки, ограничивающие отобранные исходные данные, и примененную к ним обработку. Кроме того, каждый pck-файл дополнен одним индексным ndx-файлом. Ndx-файл устроен аналогично pck-файлу и имеет то же имя (при расширении ndx вместо pck). Каждому значению средней температуры в pck-файле соответствует целое число (4 байта на значение), показывающее какое количество отдельных измерений вошло в расчет среднего значения для данного узла сетки. Ndx-файл носит служебный характер, он используется при генерации файлов *.pck. Тем не менее, он также может оказаться необходимым пользователю, например, при оценке достоверности рассчитанных средних значений в отдельных узлах. Данные типа PCK оптимальны при использовании в дальнейших вычислениях, поскольку представлены в виде чисел с плавающей запятой и могут быть при необходимости дополнены индексными файлами, характеризующими качество осреднения.

Данные в формате RAW устроены аналогично данным в формате PCK. Отличием их является то, что расчетные значения радиояркостных температур помножены на масштабирующий коэффициент, округлены до целых и записаны в целочисленном формате (один или два байта на значение). Эти данные оптимальны для непосредственной визуализации осредненных радиотепловых полей (такие raw-файлы могут быть просмотрены с использованием популярных программных продуктов, например, Adobe Photoshop) и также удобны для экспресс-анализа отобранной информации и для дальнейших расчетов.

Развитые программные средства работы с накопленными данными

Сформулированная выше идеология построения базы данных потребовала разработки специальных программных средств, обеспечивающих последовательный выбор области данных и представления выбранных данных как адаптивных (с точки зрения последующей обработки) рядов измерений. Необходимо иметь универсальное средство, позволяющее обратиться сразу к выделенному в процессе поиска в индексной базе данных классу (множеству) файлов, содержащих данные дистанционного зондирования в рамках одного сеанса обработки. Важно, чтобы функциональность созданного средства была легко расширяема для удовлетворения новых требований к выходным данным, возникающих в ходе работы с базой данных. Такое средство реализуется в форме специального драйвера базы данных.

Разработанный и реализованный драйвер базы данных состоит из двух частей: библиотеки специализированных модулей, каждый из которых содержит описание определенной операции, применимой к файлам базы данных или файлам производных типов данных, и универсального модуля, организующего «конвейерную» обработку выделенного класса файлов данных с применением одной из библиотечных операций.

Универсальный модуль («конвейер») получил название Stamper – штамповщик. Он является приложением Windows и имеет простой графический пользовательский интерфейс в виде диалогового окна, позволяющего оператору загрузить один из библиотечных модулей, указать корневую директорию обработки, определить класс обрабатываемых файлов (в виде стандартного шаблона типа “*.gz”, “myfile.dat” и т.п.) и запустить настроенный таким образом «конвейер». Задача программы Stamper – построить дерево обработки (локализовать и включить в список обработки все файлы, удовлетворяющие заданному шаблону и находящиеся в корневой директории обработки и ее поддиректориях), извлечь из загруженного библиотечного модуля специфические инструкции по обработке, применить эти инструкции ко всему списку найденных файлов. Таким образом, Stamper эффективно автоматизирует обработку длинных рядов файлов, не ограничивая ее возможностей: произвольный алгоритм обработки, описанный по определенным стандартам, будет загружен и применен к конкретному файлу или произвольному набору файлов по выбору оператора базы данных.

Библиотечные модули обработки являются динамическими библиотеками Windows специального вида. Для отличия их от прочих динамических библиотек (системных и поставляемых в составе различного программного обеспечения), они имеют файловое расширение «stp» (от *stamp* — штамп). Специфичность этих модулей заключается в обязательном наличии «входной» функции обработки, имеющей установленное имя и синтаксис. В обязанности модуля, помимо собственно обработки, входит также проверка содержимого файла на соответствие установленным критериям обработки. Так, модулю, осуществляющему поиск дистанционных данных по заданному району в заданный интервал времени, необходимо проверить фактически содержащиеся в файле данные на соответствие этим ограничениям. Настройки модуля (специфические для данной операции критерии отбора данных) хранятся в автоматически создаваемых разделах системного реестра. При необходимости, оператор имеет возможность редактировать записи реестра непосредственно, с помощью приложения regedit, находящегося в системной директории Windows. Однако для существующих библиотечных модулей созданы специальные приложения, позволяющие редактировать эти записи в диалоговом режиме при помощи простого графического пользовательского интерфейса.

Совокупность модулей обработки составляет библиотеку модулей, которая может быть неограниченно расширена путем создания новых динамических библиотек, удовлетворяющих фиксированным стандартам. Среди развитых и внедренных к настоящему моменту в ИКИ РАН модулей следует назвать Picker.stp и Mapper.stp, генерирующие выходные данные в форматах PCK и RAW соответственно.

Перспективные разработки

Практика работы с создаваемой базой данных в ИКИ РАН расширяет и уточняет набор требований, предъявляемых как к характеристикам генерируемых выходных данных, так и к методам их обработки. Среди них в качестве наиболее актуальных можно выделить следующие:

- введение типа данных аналогичных PCK или RAW, но снабженных метками времени, характеризующими момент выполнения съемки;
- расширение функциональности Picker для синтеза выходных данных по информации нескольких спутников с разной логикой объединения/дополнения;
- расширение приложения Stamper возможностью его программирования оператором с целью автоматической загрузки и последовательного исполнения нескольких библиотечных модулей обработки;
- наращивание библиотеки модулей для тематической обработки радиометрических данных, подготовки и выполнения резервного копирования информации.

Заключение

Задачи исследования вариаций климата Земли требуют интенсивного привлечения данных спутникового дистанционного зондирования, в частности, регулярных, длительных, глобальных СВЧ-радиометрических наблюдений системы океан – атмосфера. Создаваемая в ИКИ РАН база данных измерений спутникового прибора SSM/I призвана обеспечить такую возможность и организована по принципу интерпретации данных как рядов длительных глобальных измерений. Такие длительные ряды строятся из исходных данных по задаваемым пользователем критериям отбора и являются, с точки зрения пользователя, основной структурной единицей базы данных. К настоящему моменту накопленную информацию в ос-

новном составляют данные непрерывных измерений в период 1995–2003 гг., выполненных на аппаратах F10–F15 серии DMSP. Общий объем информации — около 150 Гбайт.

Для реализации заложенного в основу базы данных принципа генерации рядов длительных глобальных измерений разработано и частично внедрено специальное программное обеспечение, сочетающее эффективность универсального подхода к поиску и сбору исходных данных с гибкостью применяемых операций обработки и возможностью неограниченного расширения библиотеки таких операций.

Создаваемая база данных уже оказалась востребованной в ряде решаемых в ИКИ РАН задач, что позволило уточнить и расширить перечень требований как к форматам предоставляемых выходных данных, так и к развиваемому программному обеспечению.